technische universität
dortmund

fakultät
statistik

Modeling Approaches for Dose-Response Data in Toxicology

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund

Vorgelegt von

Julia Christin Duda

Dortmund, Januar 2024

Amtierender Dekan:
Prof. Dr. Philipp Doebler

Gutachter:
Prof. Dr. Jörg Rahnenführer (Technische Universität Dortmund)
JProf. Dr. Kirsten Schorning (Technische Universität Dortmund)
Prof. Dr. Jan G. Hengstler (Leibnizinstitut für Arbeitsforschung an der Technischen Universität Dortmund)

Tag der Prüfung:
14. März 2024

Abstract

Dose-response modeling occurs in many application areas and has a rich research history. An extensively studied application field is clinical studies, where dose-response modeling is used in Phase II studies to identify the dose closest to a pre-defined effect. Many non-clinical, toxicological studies also aim at identifying a dose-response relationship. However, for non-clinical or toxicological studies there are fewer regulations or guidelines. This leads to a gap between nowadays research advances in statistical modeling and the use of these methods in practice in toxicology. In addition, toxicological dose-response studies differ from clinical studies in various technical aspects. For example, cells might be studied instead of human patients, and administered doses are constrained due to laboratory, and technical reasons rather than ethical considerations. Therefore, the transfer of clinical methodological knowledge into toxicological applications is only possible to a limited extent and tailored methodologies are required that match the specific data structure of toxicological studies.

This cumulative thesis is based upon four works that all present approaches for modeling toxicological dose-response data. The first manuscript reveals the potential of applying the Multiple Comparison Testing and Modeling (MCP-Mod) approach by Bretz et al. (2005) developed for Phase II clinical studies on toxicological, gene-expression dose-response data. In the second manuscript, a parametric, mechanistically motivated model for toxicological dose-time-response data is developed. The third manuscript is application-focused and explains the use of interaction effects when analyzing dose-response gene expression in a two-factor setting. At last, a non-parametric Bayesian dose-response modeling approach was developed that performs functional shrinkage for non-linear function spaces. While the first three manuscripts are published, the fourth work is attached in its current version.

Acknowledgments

I would like to thank my supervisor, Jörg, for his unwavering support and optimism throughout each phase of my dissertation. Thank you for always showing me another perspective when I needed it, while also granting me a lot of freedom to explore and find my ways.

I am grateful for the opportunity to work on my Ph.D. within the Research Training Group 2624 *Statistical Methods for High-Dimensional Data in Toxicology*. It was a great experience to work on so many interdisciplinary projects. Special thanks go out to Prof. Hengstler, whose passion for his research is beyond scales and I am grateful for the opportunity to collaborate with him - thank you! I would also like to thank all members of the RTG and other colleagues, for being a great community to gain experiences, feel supported, and share fun times throughout the last years.

A special thanks goes out to Matt, the co-author of my last work, who gave me the opportunity to visit him at the National Institute of Environmental Health Sciences in the United States to work on our research project. Thank you for pushing me to work through challenges that I would have thought invincible before.

I would like to thank Ezekiel King, my roommate during my USA stay. Thank you for teaching me to become a braver version of myself and to always cherish the little things in life. I hope I made your life a little brighter in the short time we shared and that you can rest peacefully now.

At last, I would like to thank my family and friends who greatly supported me throughout all the ups and downs of this academic journey. Without your love and support, I would not have made it through the hard times.

> *"Just keep swimming."*
>
> – Dory, *Finding Nemo*.

List of Publications

This cumulative thesis is based on the following four manuscripts:

Article 1: Duda, J. C., Kappenberg, F., & Rahnenführer, J. (2022). Model selection characteristics when using MCP-Mod for dose–response gene expression data. Biometrical Journal, 64(5), 883–897. https://doi.org/10.1002/bimj.202000250

Contribution of the author:

The author of this thesis implemented the data analyses and the simulation studies, added additional ideas and wrote the first draft of the manuscript. Discussions and revising was done together with Franziska Kappenberg under the supervision of Jörg Rahnenführer.

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Article 2: Duda, J. C., Hengstler, J. G., & Rahnenführer, J. (2022). td2pLL: an intuitive time-dose-response model for cytotoxicity data with varying exposure durations. Computational Toxicology, 43, Article 100234.
https://doi.org/10.1016/j.comtox.2022.100234

Contribution of the author:

The author of this thesis had the central idea for the suggested model, implemented the software package, simulation study, and data application, and wrote the first draft of the manuscript. Helpful comments by the co-authors were incorporated.

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Article 3: Duda, J. C., Drenda, C., Kästel, H., Rahnenführer, J., & Kappenberg, F. (2023). Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons. Scientific Reports, 13(1), 20804.
https://doi.org/10.1038/s41598-023-47057-0

Contribution of the author:

The author of this thesis implemented the analyses together with Carolin Drenda and Hue Kästel, and wrote the manuscript. Comments from fruitful discussions with Jörg Rahnenführer and Franziska Kappenberg were incorporated.

Article 4: Duda, J. C. & Wheeler, M. Bayesian non-linear subspace shrinkage using horseshoe priors. Unpublished.

Contribution of the author:

The original idea for this manuscript was proposed by Matthew Wheeler. The computational implementation, design, conduction, and analysis of the simulation study and the search for and analysis of the application example as well as writing the manuscript was done by the author of this thesis under the supervision of Matthew Wheeler. The manuscript is attached in its current version.

Further publications:

1. Tug, T., **Duda, J. C.**, Menssen, M., Bruce, S. W., Bringezu, F., Dammann, M., Frötschl, R., Harm, V., Ickstadt, K., Igl, B. W., Jarzombek, M., Kellner, R., Lott, J., Pfuhler, S., Plappert-Helbig, U., Rahnenführer, J., Schulz, M., Vaas, L., Vasquez, M., Ziegler, V., Ziemann, C. (2024). In vivo alkaline comet assay: Statistical considerations on historical negative and positive control data. Regulatory Toxicology and Pharmacology, 105583. https://doi.org/10.1016/j.yrtph.2024.105583

2. Ghallab, A., González, D., Strängberg, E., Hofmann, U., Myllys, M., Hassan, R., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., **Duda, J. C.**, Drenda, C., Kappenberg, F., Reinders, J., Friebel, A., Vucur, M., Turajski, M., Seddek, A., Abbas, T., Abdelmageed, N., Morad, S. A. F., Morad, W. , Hamdy, A., Albrecht, W., Kittana, N., Assali, M., Vartak, N., van Thriel, C., Sous, A., Nell, P., Villar-Fernandez, M., Cadenas, C., Genc, E., Marchan, R., Luedde, T., Åkerblad, P., Mattsson, J., Marschall, H., Hoehme, S., Stirnimann, G., Schwab, M., Boor, P., Amann, K., Schmitz, J., Bräsen, J. H., Rahnenführer, J., Edlund, K., Karpen, S.J., Simbrunner, B., Reiberger, T., Mandorfer, M., Trauner, M., Dawson, P. A., Lindström, E., Hengstler, J. G. (2023). Inhibition of the Renal Apical Sodium Dependent Bile Acid Transporter Prevents Cholemic Nephropathy in Mice with Obstructive Cholestasis. Journal of Hepatology. Published. https://doi.org/10.1016/j.jhep.2023.10.035

3. Kappenberg, F., **Duda, J. C.**, Schürmeyer, L., Gül, O., Brecklinghaus, T., Hengstler, J. G., Schorning, K., & Rahnenführer, J. (2023). Guidance for statistical design and analysis of toxicological dose–response experiments, based on a comprehensive literature review. Archives of Toxicology. In press. https://doi.org/10.1007/s00204-023-03561-w

4. Su, H., Haque, M., Becker, S., Edlund, K., **Duda, J. C.**, Wang, Q., Reißing, J., Marschall, H., Candels, L. S., Mohamed, M., Sjöland, W., Liao, L., Drexler, S. A., Strowig, T., Rahnenführer, J., Hengstler, J. G., Hatting, M., Trautwein, C. (2023). Long-term hypercaloric diet exacerbates metabolic liver disease in PNPLA3 I148M animals. Liver International. In press. https://doi.org/10.1111/liv.15587

5. Brecklinghaus, T., Albrecht, W., **Duda, J. C.**, Kappenberg, F., Gründler, L., Edlund, K., Marchan, R., Ghallab, A., Cadenas, C., Rieck, A., Vartak, N., Tolosa, L., Castell, J. V., Gardner, I., Halilbasic, E., Trauner, M., Ullrich, A., Zeigerer, A., Turgunbayer, Ö. D., Damm, G., Seehofer, D., Rahnenführer, J., Hengstler, J. G. (2022). In vitro/in silico prediction of drug induced steatosis in relation to oral doses and blood concentrations by the Nile Red assay. Toxicology Letters, 368,

33–46. https://doi.org/10.1016/j.toxlet.2022.08.006

6. Brecklinghaus, T., Albrecht, W., Kappenberg, F., **Duda, J. C.**, Vartak, N., Edlund, K., Marchan, R., Ghallab, A., Cadenas, C., Günther, G., Leist, M., Zhang, M., Gardner, I., Reinders, J., Russel, F. GM., Foster, A. J., Williams, D. P., Damle-Vartak, A., Grandits, M., Ecker, G., Kittana, N., Rahnenführer, J., Hengstler, J. G. (2022). The hepatocyte export carrier inhibition assay improves the separation of hepatotoxic from non-hepatotoxic compounds. Chemico-Biological Interactions, 351, Article 109728. https://doi.org/10.1016/j.cbi.2021.109728

7. Brecklinghaus, T., Albrecht, W., Kappenberg, F., **Duda, J. C.**, Zhang, M., Gardner, I., Marchan, R., Ghallab, A., Turgunbayer, Ö. D., Rahnenführer, J., Hengstler, J. G. (2022). Influence of bile acids on the cytotoxicity of chemicals in cultivated human hepatocytes. Toxicology in Vitro, 81, Article 105344. https://doi.org/10.1016/j.tiv.2022.105344

8. Ghallab, A., Myllys, M., Friebel, A., **Duda, J. C.**, Edlund, K., Halibasic, E., Vucur, M., Hobloss, Z., Brackhagen, L., Begher-Tibbe, B., Hassan, R., Burke, M., Genç, E., Frohwein, L. J., Hofmann, U., Holland, C. H., González Leiva, D. F., Keller, M., Seddek, A., Abbas, T., Mohammed, E. S. I., Teufel, A., Itzel, T., Metzler, S., Marchan, R., Cadenas, C., Watzl, C., Nitsche, M. A., Kappenberg, K., Luedde, T., Longerich, T., Rahnenführer, J., Hoehme, S., Trauner, M., Hengstler, J. G. (2021). Spatio-temporal multiscale analysis of Western diet-fed mice reveals a translationally relevant sequence of events during NAFLD progression. Cells, 10(10), Article 2516. https://doi.org/10.3390/cells10102516

Contents

Part I

Introduction

1 Motivation

Dose-response analysis occurs in a variety of fields but is most intensely studied within the pharmaceutical development context. When developing a new therapeutic compound, the Phase I and II clinical trials specifically aim at finding a ’good’ dose (Ting et al., 2017). Therefore, it is crucial to understand and statistically model the dose-response relationship of the compound under investigation. The analysis of dose-response relationships is also a central goal in related fields, such as pre-clinical studies or toxicological research outside the pharmaceutical sector (Hothorn, 2016). The goal of modeling a dose-response relationship in whichever application field appears identical from a mathematical perspective, at least at first glance. But more specialized, tailored methodological approaches to each field are required due to different experimental conditions. For example, the challenge of small sample sizes in toxicological experiments calls for a different methodological development. However, the methodological research states and practices are more advanced in clinical research than in non-clinical and toxicological research for historical, political, and economic reasons. The pharmaceutical industry faces complex regulations, initiated by the Contergan scandal in the 1960s, while simultaneously being an economically very profitable sector (Bauschke, 2010). This combination quickly pushed methodological research for clinical studies towards new, complex methods that save study duration time while remaining in line with regulatory standards.

For toxicological, non-clinical studies, there are many guidelines to standardize laboratory procedures of assays and study types, for example by the Organization of Economic Co-Operation and Development (OECD) (Gordon, 2001). Laboratory testing results related to chemical safety that adhere to OECD Testing Guidelines are accepted by OECD countries. This harmonization effort reduces repeated testing for country-specific regulations and creates trust and predictability in society and the international market. Other authorities such as the European Food Safety Agency (EFSA) of the European Union or the U.S. Environmental Protection Agency (EPA) also influence policy making by providing scientific advice. Such guidelines are updated regularly to keep pace with scientific advances. Especially within the risk assessment context, there were advances in the recommendation for the statistical evaluation of dose-response data. For example, in 2009 the Scientific Committee of the EFSA published a guidance document that promoted actual modeling of dose-response relationships rather than multiple testing approaches between dose levels to determine the critical effect of a substance (EFSA, 2009).

Despite advances and efforts through guidelines and recommendations, their impact is limited as they are not binding for all toxicological studies or experiments. International guidelines apply for specific studies conducted by international companies and independent risk assessors. The vast majority of research in toxicology is not required to be guideline-coherent. This freedom empowers exploratory research but also reinforces the gap between statistical methodological research and its application in practice.

The discrepancy between widely accepted methodological advances in risk assessment and their practical use is quantified in recent a review by Kappenberg et al. (2023). From 3269 analyzed dose-response curves where some form of modeling was applied, for the vast majority (>2,250) simple linear interpolation was used. Only for about 500 curves, an actual parametric model was fitted.

In clinical research, dose-response modeling methods typically aim at dose-finding goals, i.e. finding a clinically relevant dose (Bretz et al., 2008). Mathematically, the goal is similar to finding toxicity levels, when analyzing dose-response data in risk assessment. However, methodological research for clinical dose-finding studies can be considered more prevalent. It started with the traditional rule-based, "up-and-down" 3+3 design developed in 1946 (Dixon and Mood, 1946) which was introduced for clinical trials in 1989 (Storer, 1989). Model-based designs followed such as the continual reassessment method (CRM) (O'Quigley and Shen, 1996), escalation with overdose control (EWOC) (Babb et al., 1998; Rogatko et al., 2005) and extensions such as the time-to-event CRM (TITE-CRM) (Cheung and Chappell, 2000). A hybrid approach that combines classical multiple comparison procedures (MCP) testing approaches (Tamhane et al., 1996) with modeling approaches (Pinheiro et al., 2006) is MCP-Mod (Bretz et al., 2005). Further methods and reviews are compared and referenced by Ananthakrishnan et al. (2017).

Despite this active, fast-paced development, and agreement on the superiority of the advanced methods, also in clinical trials there is a high reluctance to adopt new designs in practice (Kurzrock et al., 2021). A review by Rogatko et al. (2007) revealed that from 1,235 clinical Phase I studies published between 1991 and 2006, only 20 (1.6%) used a statistical modeling method that was published during that time period. The vast majority of the remaining studies used some form of the classical rule-based design.

The above considerations reveal how despite efforts both from academia and regulatory bodies and a theoretically very advanced research state on dose-response modeling, the application of said methods falls short in practice. Before linking how the work of this thesis contributes to filling this gap, one additional development that is intertwined with this topic is elucidated.

Another fast-developing field that effects dose-response modeling in toxicology is genomics. Starting with the discovery of the molecular structure of DNA by Watson and Crick (Watson and Crick, 1953), sequencing technologies that transform genetical information into machine-readable data were continuously improved. A breakthrough was the first complete sequence of the human genome in 2001 (Venter et al., 2001). This was achieved with the first sequencing technology, Sanger sequencing (Sanger and Coulson, 1975). The sequencing technology was rapidly advances, bringing forward microarray technology (Schena et al., 1995), next generation sequencing, which allowed massive parallel sequencing (Margulies et al., 2005), and more recently single-cell sequencing (Aldridge and Teichmann, 2020), with its first experiment in 2008. Sanger and next-generation sequencing can nowadays be considered as first and second generation sequencing. In 2009, single-cell sequencing or third generation sequencing was first introduced commercially by Pacific Biosciences (PacBio) as Single Molecule Real Time (SMRT) Sequencing (Eid et al., 2009). Third generation sequencing differs from first and second generation sequencing, as it allows much longer sequencing lengths up to tens of kilobases of DNA base pairs of a single DNA molecule. It does not require DNA amplification (Satam et al., 2023).

With next generation sequencing, the size of generated data grew rapidly. This called for new methods and software and highly influenced the research field of bioinformatics. Data generated by sequencing machines has to be pre-processed prior to the actual data analysis (Gondane and Itkonen, 2023). After pre-processing the raw reads, they have to be aligned or assembled using alignment algorithms such as STAR (Dobin et al., 2013), TopHat (Trapnell et al., 2009) or Salmon (Patro et al., 2015). Subsequent to additional normalization steps, the statistical differential gene-expression analysis follows. This analysis aims at finding genes that are differentially expressed (have different activity levels) between two groups, hence identifying genes that might cause the condition under investigation. The most used softwares to identify differentially expressed genes are DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010).

Over the last two decades, the rapidly decreasing cost of genome sequencing allowed for larger sample sizes. From $100,000,000 per human genome sequencing in 2001, the costs dropped below $1,000 per genome in 2021 (Wetterstrand, 2021). Consequently, not only two groups might be compared, but repeated measurements at different dose- or time-points became financially feasible. Such experimental setups allowed for dose-response modeling considerations in the field of RNA sequencing data.

In summary, dose-response modeling is a central task for toxicological risk assessment and faces various, interdisciplinary challenges. It is intensively studied in related fields such as clinical trials, where the mathematical approach is similar, though contextual differences lead to different focuses. Additionally, ceaseless technical developments in genomics increase RNA sequencing data set sizes such that dose-response modeling is starting to become relevant in bioinformatical analyses, too. Despite regulations and guidelines, the transfer from state-of-the-art methodological dose-response modeling research into practice is slow both for toxicological and clinical research. Overarching use of general dose-response modeling approaches developed in the context of either clinical or toxicological research is rare but promising, due to the mathematical similar goal. Specific, contextual differences between toxicological and clinical studies certainly prevent a general interchangeability of methods. For example, censored survival data in clinical trials are typically not present in controlled toxicological studies with mice or cells, or exposure times in clinical models typically refer to the time after an administration, rather than the time period of a continuous exposure duration. Such differences do not allow indiscriminate methodological interchangeability, but uphold the need for tailored method developments.

The works of this thesis contribute to the aforementioned challenges from various perspectives. The first paper reveals the potential of using MCP-Mod - a dose-response modeling approach developed in the clinical research context - on toxicological, in vitro gene expression dose-response data (Duda et al., 2022b). MCP-Mod is a two-step procedure designed to first detect a Proof-of-Concept (PoC) of a candidate drug and then, conditioned on an established PoC, determine the dose of interest by considering model uncertainty. These ideas map well to the goals of gene expression analysis, where the detection of active genes is of interest, while the dose-response model of each gene is uncertain a priori.

The second paper introduces a time-dose-response modeling approach that is tailored to exposure-time dose-response modeling for cytotoxicity experiments (Duda et al., 2022a). Motivated by Haber's law (Haber, 1924), the model links a physiological idea to a popular dose-response model, yielding a time-dose-response model. Equipped with an R software package and a recommended two-step procedure, the approach is presented with a practical focus, thereby minimizing the burden for potential users.

Promoting the optimal use of existing tools and methods in RNA Sequencing analyses in practice is the aim of the third paper of this thesis. A common experimental scenario in RNA Sequencing is the comparison of a factor with two (or more) levels between two groups (Duda et al., 2023). Statistically, research questions for these scenarios naturally translate into the analysis of an interaction effect. However, in practice, methodological workarounds are often used. The paper explains and demonstrates in detail using an RNA-Seq mouse data set, the benefits of directly modeling interaction effects compared to the commonly used approach.

The last manuscript presents a new modeling approach that combines non-parametric modeling with functional shrinkage into a parametric space. The approach shrinks a non-parametric model into a pre-defined parametric function space if the data does not suggest deviations from that space. If the data presents deviations from the assumed function space, the approach is not limited to the assumed space and remains flexible. We demonstrate the particular use of this general method for dose-response modeling, where often the parametric Hill model is considered plausible, even though deviations (e.g. due to downturn effects at large doses) cannot be ruled out in advance and must be modeled flexibly. Further, the approach's adaptive behavior is appealing for modeling high throughput dose-response data, as it reduces the burden of correctly selecting a fixed set of candidate models that is appropriate for thousands of genes.

The remainder of this thesis is structured as follows. In Chapter 2, a methodological background is provided. It contains general concepts and methods that are used in the works of this thesis but were developed before and are not part of the research contribution of this thesis. Chapter 3 provides a summary of each of the four papers. Close attention is paid to pointing out the innovative aspect of each work and how it contributes to the respective research landscape. The thesis closes with a discussion in Chapter 4. All full-length papers are attached thereafter.

2 Statistical Methods

In this chapter, a general methodological background of existing methods that were used or extended in the works of this thesis is presented. To allow intuitive notations, symbols are mostly but not strictly used consistently across topics.

2.1 Dose-response models

For the remainder of this thesis, we use the terms dose and concentration interchangeably, despite its differences from a biological perspective. Mathematically, there is no distinction, and for brevity, we will mostly refer to a dose.

Consider $n \in \mathcal{N}$ doses $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ with corresponding response values $\boldsymbol{y} = (y_1, \dots, y_n)^\top$. Doses might be grouped into $k < n$ distinct dose levels $d_1, \dots, d_k$ with $n_1, \dots, n_k$ repeated response measurements and $\sum_{i=1}^{k} n_i = n$.

Assume a parametric relationship between the true mean response and the dose level, and additive, mean-zero, homoscedastic noise with variance $\sigma^2 > 0$:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

with $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathcal{N}(\mathbf{0}, I_n\sigma^2)$. The parametric function f is unknown, and has to be modeled. The following parametric dose-response models were considered in the papers.

The **Hill model** (a.k.a. sigmoidal Emax model or four-parameter log-logistic model) was developed by A. V. Hill in 1910 (Hill, 1910) to describe biochemical equilibrium between oxygen availability and hemoglobin saturation. It is one of the most frequently used dose-response models (Goutelle et al., 2008) and has a characteristic sigmoidal shape. The Hill model is known under different names and parameterizations (Ritz, 2010). We use the parametrization as in the R package `DoseFinding` (Bornkamp et al., 2010):

$$f_{\boldsymbol{\theta}}(x) = E_0 + E_{\max}\frac{x^h}{ED_{50}^h + x^h} \tag{2.2}$$

with parameters $\boldsymbol{\theta} = (E_0, E_{\max}, ED_{50}, h)^\top$. The intercept or background response at zero dose ($x = 0$) is E_0. $E_{\max}$ is the asymptotic maximum and ED_{50} is the dose yielding half of the asymptotic maximum effect. The parameter h defines the steepness of the curve at ED_{50}. For the special case $h = 1$, the Hill model becomes the **Emax model**, which has an asymptote only for large doses, but not at the zero dose level.

The **Beta model** is defined as

$$f_{\boldsymbol{\theta}}(x) = E_0 + E_{\max}B(\delta_1, \delta_2)(x/scal)^{\delta_1}(1 - x/scal)^{\delta_2} \tag{2.3}$$

with $B(\delta_1, \delta_2) = (\delta_1 + \delta_2)^{(\delta_1+\delta_2)}/(\delta_1^{\delta_1}\delta_2^{\delta_2})$ capturing the shape of the density function of a Beta distribution on $[0, scal]$. The Beta model can account for downturn effects after an increase and is a comparatively flexible parametric dose-response model. While the scaling parameter $scal$ is a pre-defined hyperparameter that is by default set to 1.2 times the maximal dose, the parameters to be estimated are $\boldsymbol{\theta} = (E_0, E_{\max}, \delta_1, \delta_2)^\top$. The linear parameters E_0 and $E_{\max}$ can be interpreted as for the Hill model, while $\delta_1 > 0$ and $\delta_2 > 0$ are non-linear parameters that define the location and steepness of the mode.

Another simpler non-linear model is the **Power model**

$$f_{\boldsymbol{\theta}}(x) = E_0 + E_{\max}x^h \tag{2.4}$$

with power coefficient h defining the steepness and potency in the dose-response context. It is used in the U.S. EPA's Benchmark Dose Software (BMDS) (Davis et al., 2011).

Further, simple linear regression models or quadratic models are candidate dose-response models and are assumed familiar.

In the frequentistic setup, model fitting can be performed by ordinary least squares, which coincides with a maximum likelihood approach for the normal error assumption, or weighted least squares. In the Bayesian framework, the posterior distribution is calculated analytically or approximated computationally. The specific computational implementation for both frameworks varies. For examples and details, see Pinheiro et al. (2014) for the MCP-Mod Software, Ritz et al. (2015) for the `drc` R package (frequentistic), or Wheeler et al. (2023) and Davis et al. (2011) for the bayesian `ToxicR` R package and BMDS software.

2.2 MCP-Mod

MCP-Mod was developed for Phase I/II clinical dose finding studies by Bretz et al. (2005) to analyze dose-response data and thereby combine two classical approaches: Multiple comparison procedures (MCP) and modeling (Mod). It is a dual approach to address two goals simultaneously: Performing a proof-of-concept (PoC) step, and if the PoC is established, model the dose-response relationship to derive the dose of interest. With the PoC step or MCP step, it is analyzed if the true dose-response f is non-flat, i.e. a dose-response signal exists. More precisely, one tests if f is within a set $\mathcal{M}$ of considered dose-response functions, or not. Practitioners a-priori define reasonable, standardized model shapes $f_1^0, \ldots, f_M^0$ by decomposing f_m into a location, scale and shape component:

$$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(x, \boldsymbol{\theta}^0).$$

A guesstimate approach translates interpretable statements on the expected response into quantifiable shape parameters $\boldsymbol{\theta}^0$ for each candidate model. Each model shape f_m^0 then defines a mean response vector $\boldsymbol{\mu}_m$. The PoC test is constructed based on the hypotheses $\mathrm{H}_0^m : \boldsymbol{c}_m^\top \boldsymbol{\mu}_m = 0$ and $\mathrm{H}_1^m : \boldsymbol{c}_m^\top \boldsymbol{\mu}_m \neq 0$ where the contrast vector $\boldsymbol{c}_m$ is constructed to maximize the power of the test under the assumption that $\boldsymbol{\mu}_m$ is the true mean response. Each contrast defines a test statistic

$$T_m = \frac{\sum_{i=1}^k c_{mi} \bar{y}_i}{S \sqrt{\sum_{i=1}^k c_{mi}^2 / n_i}}, \tag{2.5}$$

where $\bar{y}_i$ is the mean response at distinct dose level d_i and $S^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n-k)$ with y_{ij} being the jth measurement at dose d_i. Given the normal distribution assumption in (2.1), $(T_1, \ldots, T_m)$ follows a multivariate normal distribution. A PoC is established if $T = \max(T_1, \ldots, T_m)$ is greater than the corresponding quantile q_α, and α is the pre-selected significance level.

This multivariate approach controls the family-wise error rate while remaining powerful because correlations between similar model shapes are accounted for automatically. If a PoC is established, a non-flat dose-response relationship is assumed. All models that passed the PoC step are subject to modeling the dose-response curve in the modeling (Mod) step. In the Mod-step either model selection using the Akaike Information Criterion (AIC) or model averaging with weights based on the AIC is performed. MCP-Mod is used in the first paper of this thesis, with a focus on model selection via the AIC criterion.

2.3 Gene expression data

Gene expression data cover a wide variety of data types and technologies to generate them, and algorithms for downstream analyses. A general historical and technical overview was provided in Chapter 1. With a focus on the papers of this thesis, this section provides more details on microarray data and RNA Sequencing data. For elementary biological concepts on cells, proteins and DNA, see Brazma et al. (2001).

Microarrays quantify gene activities, by quantifying a single sample's messenger RNA (mRNA), a gene product assumed proportional to a gene's activity, at thousands of genes simultaneously. Small mRNA molecules - transcripts - are labeled with a fluorescent dye and become targets. The transcripts are deposited over the microarray chip. On the chip's surface, there are small oligonucleotides (RNA strands) called probes which the sample's targets can bind to (hybridization), if the genetic sequence is similar enough. In theory, only compatible probe targets remain on the ship after a washing phase. Fluorescence-based intensity measurements are assumed proportional to the corresponding gene activities (Sánchez and de Villa, 2008).

There are further technical details such as adjusting for background fluorescence that originates from non-specific hybridization, i.e. mRNA samples that bind to non-complementary probes. After quality control assessment, background signal adjustment and normalization to account for technical artifacts can be performed using the robust multi-array averaging (RMA) measure (Irizarry et al., 2003). The final, pre-processed intensity values of different samples are typically on a $\log_2$ scale and considered approximately normally distributed.

An important conceptual difference between microarrays and next generation, RNA-seqeuencing, is that microarrays have a pre-defined set of known target sequences, while RNA-sequencing also allows de novo assembly of unknown genomes (Gondane and Itkonen, 2023).

The bioinformatical assembling of the reads is not subject to the research of this thesis and shall not be discussed here. Starting with the resulting raw read counts, normalization and statistical modeling follow, as summarized in Chapter 1. In the papers of this thesis, the R package `DESeq2` (Love et al., 2014) is used to identify differentially expressed genes (DEGs). The relevant modeling and normalization of `DESeq2` is explained in the following (cf. Love et al. (2014)). The read counts of gene i for sample j are modeled using a negative binomial distribution, $K_{ji} \sim \text{NB}(\mu_{ij}, \alpha_i)$, where μ_{ij} is the expected mean read count and α_i a dispersion parameter. To account for different gene lengths and other technical considerations, model $\mu_{ij} = q_{ij}s_j$, where s_j is the size factor or technical read depth for sample j, and q_{ij} is the unknown, true DNA fragment concentration of gene i for sample j. The size factors s_j are estimated using the median-of-ratios method (Anders and Huber, 2010). The fragment concentration that represents the gene activity is then modeled using a generalized linear model with a logarithmic link: $\log_2 q_{ij} = \sum_{r=1}^{R} x_{jr}\beta_{ir}$, where x_{jr} dummy-codes the experimental conditions $r = 1, \dots, R$ in a design matrix $X = \{x\}_{jr}$ and β_{ij} are the coefficients. The selection of the model design and corresponding interpretation of the model fit for typical gene expression experiments is the topic of the third paper of this thesis.

2.4 Functional shrinkage for linear subspaces

For the regression problem outlined in (2.1), Shin et al. (2020) developed a non-parametric, Bayesian, flexible modeling approach that shrinks the posterior mean response towards a pre-specified linear subspace $\Omega_0 = \{\Phi_0 \boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^n\}$. The approach is adaptive because there is no shrinkage if the data suggests that the true mean response is outside of Ω_0. The mean response is modeled non-parametrically using B-splines (Carl, 2001):

$$f(x) = \sum_{m=1}^{k} \beta_m \phi_m(x),$$

where ϕ_m are B-spline bases evaluated at $m = 1, \dots, k$ knots and design matrix $\Phi = \{\phi_m(x_i)\}_{i,m}$. Shrinkage of the posterior mean response $\mathbb{E}(\Phi\beta|\boldsymbol{y})$ into Ω_0 is enforced by the conditional prior specifications

$$p(\boldsymbol{\beta}|\sigma^2, \tau^2) \propto (\tau^2)^{-(k-d_0)/2} \exp\left(-\frac{1}{2\sigma^2\tau^2}\boldsymbol{\beta}^\top \Phi^\top (I - P_{\Phi_0})\Phi\boldsymbol{\beta}\right). \tag{2.6}$$

and

$$p(\tau) \propto \frac{(\tau^2)^{b-1/2}}{(1+\tau^2)^{(a+b)}} \mathbb{1}_{(0,\infty)}(\tau), \tag{2.7}$$

where $d_0 = \text{rank}(\Phi_0)$, and $P_{\Phi_0} = \Phi_0(\Phi_0^\top \Phi_0)^{-1}\Phi_0^\top$ is the orthogonal projection matrix into the column space of Φ_0. For example, shrinkage into a simple linear subspace can be implemented using $\Phi_0 = \{\boldsymbol{1}\, \boldsymbol{x}\}$. The prior specification on τ, the shrinkage or scaling parameter, is a horseshoe, standard half Cauchy prior for $a = b = 1/2$ (Carvalho et al., 2010). It allows strong shrinkage of weak signals close to zero while leaving strong signals unconstrained. In the model, this shrinks weak deviations of the response from

Ω_0 such that the response is effectively in Ω_0, while strong deviations of the response from Ω_0 remain unconstrained. For the hyperparameters a and b, Shin et al. (2020) provides boundaries that guarantee optimally fast adaptive behavior (shrinkage or no shrinkage) w.r.t. the sample size.

In the forth paper of this thesis, the approach by Shin et al. (2020) is extended to non-linear function spaces. The relevance of this extension for toxicological applications is highlighted as it is demonstrated on the non-linear Hill model.

3 Summary of the Articles

3.1 Article 1: Model selection characteristics when using MCP-Mod for dose-response gene expression data

The first article contributes to the research area of dose-response modeling in a translational way. We identified MCP-Mod - a modeling method developed for clinical dose-response analyses - as an adequate tool for dose-response modeling in a field where MCP-Mod has not yet been used: gene expression data. Application of MCP-Mod in gene-expression dose-response modeling is motivated by uncertainty considerations. Gene-expression data are high-dimensional and it is therefore not feasible to carefully choose appropriate modeling constraints for each gene. Also, differences in cell experiments compared to human dosing data lead to further uncertainties in cell-based dose-response data. If doses are applied to cells, there are fewer ethical constraints on higher doses. This might lead to deviations from expected shapes. For example, downturn effects due to systematic toxicity at higher doses violate monotonicity assumptions. Uncertainty regarding the final dose-response shape for gene expression data is hence imminent, both due to the large number of genes that can not be viewed individually as well as less constraints on dose selection. MCP-Mod addresses the issue of model uncertainty by allowing a set of models to be considered, from which a single best model is used or models are averaged for the final fit.

We apply and analyze the model selection workflow of MCP-Mod on a gene-expression data set by Krug et al. (2013), where several doses of valproic acid (VPA) are applied to embryonic stem cells. The six models linear, quadratic, sigmoidal Emax, Emax, and beta model, are considered.

The aim of the first article was to analyze if any of the considered models are more relevant than others, or might even be superfluous for the analysis. To address this question, MCP-Mod was applied to the VPA data set, a simulation study was conducted and a score that grades the model set's performance was proposed.

The analyses lead to several conclusions that are relevant in practice and motivated the use of MCP-mod in other applied dose-response projects that face uncertainty (e.g. Brecklinghaus et al. (2022)). The monotonicity assumption should not be stated for analyzing dose-response gene-expression data. Models that were able to capture a down-turn effect (beta and quadratic) cannot be excluded from the model set without performance loss. The popular Emax model is a special case of the more general, and also very popular sigmoidal Emax model. However, the Emax model can be excluded without notable performance loss, especially if the sigmoidal Model is included in the model set. The sigmoidal model is often selected if the response signal is clear. This result aligns with the sigmoidal Emax model's popularity in practice and its mechanistically motivated origin. Often, gene expression data is noisy due to the many possible sources of variability. For such genes, the linear model is useful for the detection of a general trend, though the true dose-response shape likely is not linear and cannot be detected accurately due to the noise.

To sum up, the first article laid an important groundwork for the cumulative work of this thesis, which thereafter proceeded to focus on dose-response modeling, gene-expression data, and uncertainty considerations.

3.2 Article 2: td2pLL: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations

The second article expands the research scope of dose-response modeling to time-dose-response (TDR) modeling. TDR is often considered in the clinical context when the time after intake of a dose is considered. However, this work focuses on TDR modeling in the toxicological context where cells are continuously exposed to a dose of a compound for a certain amount of time. Cells can be exposed continuously throughout a time period in contrast to the momentary intake of a dose for humans, as cells are in a solution that contains the administered concentration.

The contribution to the TDR research of this article is the proposition of a new time-dose-response model tailored to cytotoxicity experiments and the development of an R software package to facilitate its use. The proposed model is called time-dose 2-parameter log-logistic (td2pLL) model. It is a two-dimensional, parametric approach and relates the ED_{50} parameter to the exposure time:

$$ED_{50}(t) = \Delta t^{-\gamma} + C_0, \tag{3.1}$$

which leads to the TDR model

$$f(t,x) = 100 - 100\frac{x^h}{x^h + ED_{50}(t)^h}. \tag{3.2}$$

This relationship is mechanistically motivated by a modification of Haber's law, which states that the toxic effect of a compound is the product of the concentration and the exposure time (Haber, 1924). The added parameters Δ, γ and C_0 are interpretable as the magnitude of the exposure-time dependent effect, its potency, and a potential minimal concentration toxicity threshold. The latter means that a minimal dose is required to yield a toxicity effect for a hypothetical, infinitely long exposure duration. The lower and upper asymptotes of 100 and 0 are considered fixed, as the model is tailored to cytotoxicity experiments where initially all cells live and then continuously lose viability with increasing dose.

The resulting model is more complex for practitioners than fitting a single dose-response curve for each considered exposure time. However, the model has several advantages if the experimental ranges of the dose and the exposure period allow to observe a time-dependent change in the ED_{50}. These include an increased precision in the ED_{50} estimate and the possibility to extrapolate ED_{50} for time periods not considered in the experiment.

To decide whether the benefits justify the model's complexity, a two-step pipeline was suggested. In the first step, an ANOVA-based approach is used to verify if the ED_{50} depends on the exposure duration. If there is a dependency, the td2pLL model is used in the second step. If not, this joint model is considered unnecessarily complicated, and separate dose-response curves are fitted at each exposure duration.

The two-step approach using the suggested model is compared to other approaches (always using the td2pLL model, always fitting separate time-doses curves per time duration, ignoring the time data and fitting a single time-dose curve) in a simulation study and applied on a real cytotoxicity data set by Gu et al. (2018).

The simulation results suggest that the two-step approach successfully decides between using the td2pLL model or the simpler approach and leads to more precise ED_{50} estimates. The real data demonstration also indicates the benefit of the proposed model and pipeline. However, the low number of different exposure points per compound used in the experiments limits the reliability of these conclusions. Larger data examples are required to provide a stronger empirical evaluation of the mechanistically derived ED_{50} dependency.

In summary, the suggested pipeline together with the provided software underlines the application-oriented focus of this work. The article targets a broader audience, possibly with less profound statistical knowledge, to help bridge the gap between methodological knowledge and its use in practice in toxicological risk assessment.

3.3 Article 3: Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons

The third article is an application-oriented work and was motivated by the interdisciplinary projects together with the Leibniz Research Centre for Working Environments and Human Factors (IfADo), e.g. by Su et al. (2023). It addresses how a statistically well-known, potentially beneficial method has little use in practice for a very common research situation in gene expression analyses. Precisely, we explain and illustrate the use of interaction effects in gene expression analyses of experimental situations with two factors with two levels each, or where one factor has more levels.

A typical experimental setup might compare the effect of a treatment (using Treatment A or B) on two groups of genetically modified mice (Genotype 0 and 1). The biological research question can then be formulated as: Does the treatment effect differ w.r.t. the genotype? Statistically, this naturally translates into analyzing an interaction effect γ in the model

$$y_j = \mu + \alpha g_j + \beta c_j + \gamma g_j c_j + \varepsilon_j,$$

where μ is the mean in the reference group (genotype = 0, treatment = A), $c_j = 1$ if the treatment assigned to sample j is 1, otherwise $c_j = 0$. Equivalently the genotype of sample j is indicated by g_j and ε_j is some error term. To test whether the treatment effect differs between the genotypes, one considers the hypotheses $H_0 : \gamma = 0$ and $H_1 : \gamma \neq 0$.

The above approach is rather simple. However, while for many experimental situations in practice, it would be appropriate to analyze them using the above approach, the reality is different. Often a statistical detour is used that aims at analyzing interactions, but evades the concept of (statistical) interaction effects in a model. Precisely, practitioners might split the data set into samples of genotype A and samples of genotype B. Then, separately for each genotype data set, the treatment effect is estimated. If the treatment effect is significant (and possibly relevant) for one genotype data set, but not for the other, a difference in the treatment effects w.r.t. the genotype is concluded. We denote this commonly found approach as ’Method I’ and the interaction effect-based approach as ’Method II’.

In the article, we explain and demonstrate the differences and potential benefits of Method II - modeling an interaction effect - using a gene expression data set where mice are fed two different diets over many weeks. The diet was either a healthy standard diet (SD) or an unhealthy, fatty western diet (WD) and we focused on the feeding periods of either 3 weeks or 6 weeks. The effect of the WD over the SD is of interest. Biologically, genes where the diet effect is different after 6 weeks of feeding compared to the diet effect at the reference time of three weeks of feeding are of interest. Statistically, this calls for modeling an interaction effect between diet type and feeding period. As we dealt with RNA sequencing data, the expression levels are modeled using a Negative Binomial distribution and a generalized linear model, as described in Section 2.3, and implemented using the `DESeq2` R package (Love et al., 2014). Method I and method II were compared on the data set and results suggest that modeling an interaction effect can lead to smaller but more specific gene sets.

The article provides practitioner-orientated explanations and real-data demonstrations on using interaction effects tailored to gene expression analyses. Gene expression analyses are a standard experimental procedure used by many laboratories with different levels of statistical training available in-house. Therefore, the article contributes to bridging the gap between available statistical research and software for gene expression analyses and its optimal use in practice.

3.4 Article 4: Bayesian non-linear subspace shrinkage using horseshoe priors

This article extends the work of Shin et al. (2020), who introduced shrinkage of non-parametrically modeled response curves into parametric linear subspaces, towards shrinkage into non-linear function spaces (cf. Section 2.4). In contrast to the previous articles, this work is formulated within a Bayesian framework. While the work’s extension towards non-linear function spaces is generally applicable for any non-linear parametric model, its particular relevance for toxicological research is demonstrated by shrinkage into the non-linear Hill model (cf. Section 2.1). The underlying motivation for this work is that mechanistically motivated models as the Hill model are often non-linear and always limited to being an approximation of a more complex, underlying physiological mechanism. Therefore, deviations from a model that is overall plausible, cannot be ruled out beforehand and a flexible modeling approach is needed that shrinks into the reasonable model, but is not fully constrained to the model space.

The methodological extension is based on conditional linearization of the non-linear function space using a first-order Taylor expansion. Let h_θ be the Hill model function (cf. 2.1) and $\Omega_0^{\boldsymbol{\Theta}} = \{h_{\boldsymbol{\theta}}(x) | \boldsymbol{\theta} \in \boldsymbol{\Theta}, x \in \mathbb{R}_0^+\}$ the corresponding function space. $\Omega_0^{\boldsymbol{\Theta}}$ cannot be represented by a matrix column space as for linear subspace shrinkage. Therefore, consider a linear approximation of $\Omega_0^{\boldsymbol{\Theta}}$ at $\boldsymbol{\theta}_0$ for fixed $\boldsymbol{x}$:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx h_{\boldsymbol{\theta}_0}(\boldsymbol{x}) + \dot{F}_{\boldsymbol{\theta}_0}(\boldsymbol{x})(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{3.3}$$

with partial derivatives $\dot{F}_{\theta_0} = \dot{F}_{\theta_0}(\boldsymbol{x}) = \left.\frac{\partial h_{\boldsymbol{\theta}}(x)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \in \mathbb{R}^{n \times s}$. Instead of P_{Φ_0} for the linear case, we propose to use $P_{\boldsymbol{\theta}} = P_{\dot{F}_{\boldsymbol{\theta}}}$ as projection into the column space of $\dot{F}_{\boldsymbol{\theta}}$ for a given $\boldsymbol{\theta}$. Geometrically, this can be justified as $\dot{F}_\theta(x)$ locally approximates $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ by spanning tangent planes at each $x_i \in \boldsymbol{x}$ (Seber and Wild, 2003)[p. 130]. The resulting, conditional shrinkage prior for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\sigma^2, \tau^2, \boldsymbol{\theta}) \propto (\tau^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2\tau^2}\boldsymbol{\beta}^\top \Phi^\top (I - P_{\boldsymbol{\theta}})\Phi\boldsymbol{\beta}\right). \tag{3.4}$$

Notably, the prior now depends on $\boldsymbol{\theta}$, which has to be updated and provided with prior distributions as opposed to the linear shrinkage of Shin et al. (2020). The prior choices for $\boldsymbol{\theta}$ depend on the choice of $h_{\boldsymbol{\theta}}$ and available knowledge on the expected response. Further, the conditional precision matrix of $\boldsymbol{\beta}$ has full rank due to the non-linearity of $h_{\boldsymbol{\theta}}(x)$, leading to $(\tau^2)^{-k/2}$ instead of $(\tau^2)^{-(k-d_0)/2}$ in the linear shrinkage case.

The proposed method can further be extended to shrinkage into combined function spaces. Therefore, the corresponding $\dot{F}_\theta$ of different $h_{\boldsymbol{\theta}}$ are combined horizontally to form an overall derivative matrix, under the constraint that only one intercept column remains. This straightforward extension to shrinkage into combined function spaces allows for additional robustness and can properly capture uncertainty if more than one parametric model is reasonable.

We demonstrate the method for shrinkage into the single Hill model space and the combined space of the Hill model and the Power model. In an extensive simulation study, these method implementations are compared to various alternative approaches, namely B-splines, P-splines, a parametric fit with an additive horse-shoe shrinkage spline, and common parametric fits (correctly or incorrectly specified). Additionally, the method is applied to real data where age-dependent testosterone levels in men are modeled.

As a result, the non-parametric non-linear functional shrinkage approach performs comparative to the oracle-condition parametric fit in case of correct function space specification. If the wrong function space is specified, it correctly decides against shrinkage into the function space and effectively becomes an unconstrained B-spline fit as by construction. Shrinkage into combined function spaces leads to better fits due to the increased robustness. In the context of dose-response modeling, this feature is helpful for the common scenario of down-turn effects at higher doses. Here, the response curve matches a Hill model up to a certain dose where a downturn starts, for example, due to systematic toxicity. A limitation of the proposed method is that in the case of function space misspecification, the missing shrinkage leads to too unconstrained fits that are prone to overfitting. Approaches to overcome this and other limitations are discussed in Section 4.

4 Discussion and Outlook

For this thesis, four manuscripts with different forms of contributions to the dose-response modeling research landscape in toxicology were presented. Dose-response modeling is a large field with major applications in clinical trials, risk assessment, or toxicology. Its methodological developments are additionally influenced by the fast technological developments in high-throughput testing and genomics.

The first manuscript applied the dose-response modeling method MCP-Mod, originally developed within the clinical context, for high-dimensional, toxicological dose-response gene-expression data (Duda et al., 2022b). This application is innovative as the transfer between dose-response modeling approaches between clinical applications and toxicological applications is rare, despite large methodological overlaps of the two fields. Here, the identification of the methodological transfer potential has an additional novelty aspect as the transfer is not only from the clinical towards the general toxicological context but to more specific, high-dimensional, toxicological gene expression data. The main motivation to use MCP-Mod for gene-expression dose-response data is the method's ability to handle model uncertainty, which is crucial when modeling thousands of genes simultaneously. As a result of this work, the necessity of considering model uncertainty in gene-expression data was demonstrated. To quantify this, tailored measures and evaluation procedures were additionally developed and a large simulation study was performed.

The first paper's scope was on model *selection* using MCP-Mod and further investigations on dose-response model *averaging* leave opportunities for future investigations. Benchmark dose (BMD) estimation based on model averaging is a promising approach. However, the inclusion of down-turn effect models might undermine the desirable plateau at low doses and require further investigation. Further, multivariate approaches to combine BMDs, potentially w.r.t. functional genetic groups, but within limits of computational feasibility, are research opportunities based on this work. Related to this task are the works of Kappenberg and Rahnenführer (2023) and the BMDExpress 2 software of the National Toxicology Program (Phillips et al., 2019). Kappenberg and Rahnenführer (2023) uses an empirical Bayes framework to share Hill model parameter information in functional genetical groups to improve the estimation of gene-wise and gene-group-wise effective doses. The work is limited to the Hill model as a basis for the parameter-sharing approach. BMDExpress 2 considers various models and genetic pathway information to pool BMDs of genes within functional groups. The pooling is limited to simple summary statistics, and challenges to incorporate fast, multivariate approaches remain.

The second paper of this thesis proposed a mechanistically motivated two-dimensional time-dose-response (TDR) model tailored to cytotoxicity experiments. Additionally, a two-step guideline for parsimonious use of the model was provided and the methodology is made available in a developed R-package (Duda et al., 2022a). The work is an example that many toxicological applications benefit from tailored dose-response or TDR method development, demonstrating the limitations of methodological interchangeability between clinical and toxicological dose-response research. A limitation of the proposed model is the small scale of the data used for the application, which limits statements on empirical validation of the model. Larger, appropriate cytotoxicity data sets would also promote the proposition of more parametric TDR models, opening doors for model averaging considerations in the TDR modeling context.

The third paper of this thesis is application-focused and aims at promoting the optimal use of available methods and software when analyzing RNA-sequencing data in dose-response two-group comparisons. In particular, correct modeling of an interaction effect, its implementation in the R-package `DESeq2` (Love et al., 2014), thorough explanations, and a demonstration of its methodological benefits on an RNA-sequencing mice data set are provided. To build upon this work, extensive simulation studies and literature search-based work that empirically assesses the potential of modeling interaction effects on conducted studies offer a research outlook.

The last work of this thesis proposes a general Bayesian non-linear functional shrinkage (NLFS) approach that can specifically be applied in the dose-response context. Based upon a non-parametric model, the response adaptively shrinks towards a pre-defined, non-linear function space (e.g. the Hill model), or a combination of function spaces. NLFS is adaptive as shrinkage is not enforced if the data suggests that the function space is misspecified, yielding a highly flexible approach that accounts for prior knowledge, but also allows deviations. The method is an extension of Shin et al. (2020), who developed this method for linear subspace shrinkage. The extension to non-linear function spaces is based on a first-order Taylor expansion of the model space of interest. In an extensive simulation study, NLFS was compared to other modeling approaches and applied to a testosterone data set. To speed up computation, a tailored block Gibb's sampler is implemented, which uses Metropolis-Hastings and Slice Sampling.

The simulation results suggest that the proposed non-parametric method successfully approximates a parametric fit if the function space is correctly specified. If the function space is misspecified, the shrinkage quickly (for low sample sizes) vanishes, allowing an unconstrained fit. We also demonstrated that shrinkage towards a combination of approximated function spaces is straightforward and provides additional robustness as further model uncertainty considerations can be met. An application example of testosterone levels modeled for men of any age highlights the method's strengths, as the response is similar to the Hill model, while plausible, data-driven deviations from the Hill model are present.

Several improvements and extensions of the proposed NLFS approach offer further research directions. By construction, the model is a flexible non-parametric, unconstrained B-spline if the assumed function space is severely missspecified. This can lead to undesirable overfitting. To overcome this limitation, the implementation of a smoothness penalty as suggested by Wiemann and Kneib (2021) appears promising. Further, the NLFS approach cannot handle a few ($n < 10$), distinct dose levels with repeated measurements, which is common in practice. So far NLFS is implemented based on uniformly drawn doses, which guarantees balanced shrinkage across the dose range and circumvents rank-deficiency challenges. Adapting the approach and implementing an additional grid that regulates the shrinkage independent of the doses is worth further investigations. Within this extension, additional weights that allow for less shrinkage at the end, e.g. to better account for down-turn effects, would meet further practical needs. However, a grid approach requires additional hyperparameter considerations for the grid size and density. To avoid these limitations, embedding the approach into a Gaussian Process framework is another research perspective.

Bibliography

Aldridge, S. and Teichmann, S. A. (2020). Single cell transcriptomics comes of age. *Nature Communications*, 11(1):4307.

Ananthakrishnan, R., Green, S., Chang, M., Doros, G., Massaro, J., and LaValley, M. (2017). Systematic comparison of the statistical operating characteristics of various phase i oncology designs. *Contemporary clinical trials communications*, 5:34–48.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17(10):1103–1120.

Bauschke, R. (2010). *The Effectiveness of European Regulatory Governance: The Case of Pharmaceutical Regulation*. PhD thesis.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2010). Dosefinding: planning and analyzing dose finding experiments. *R package version 0.4-1*.

Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology-cells, molecules, genes, functional genomics, microarrays. *European Bioinformatics Institute, Draft*.

Brecklinghaus, T., Albrecht, W., Kappenberg, F., Duda, J., Vartak, N., Edlund, K., Marchan, R., Ghallab, A., Cadenas, C., Günther, G., et al. (2022). The hepatocyte export carrier inhibition assay improves the separation of hepatotoxic from non-hepatotoxic compounds. *Chemico-biological interactions*, 351:109728.

Bretz, F., Hsu, J., Pinheiro, J., and Liu, Y. (2008). Dose finding–a challenge in statistics. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4):480–504.

Bretz, F., Pinheiro, J. C., and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3):738–748.

Carl, D. (2001). *A practical guide to splines*. Springer.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Cheung, Y. K. and Chappell, R. (2000). Sequential designs for phase i clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182.

Davis, J. A., Gift, J. S., and Zhao, Q. J. (2011). Introduction to benchmark dose methods and us epa's benchmark dose software (bmds) version 2.1. 1. *Toxicology and applied pharmacology*, 254(2):181–191.

Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

Duda, J., Hengstler, J. G., and Rahnenführer, J. (2022a). td2pll: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations. *Computational Toxicology*, 23:100234.

Duda, J. C., Drenda, C., Kästel, H., Rahnenführer, J., and Kappenberg, F. (2023). Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons. *Scientific Reports*, 13(1):20804.

Duda, J. C., Kappenberg, F., and Rahnenführer, J. (2022b). Model selection characteristics when using mcp-mod for dose–response gene expression data. *Biometrical Journal*, 64(5):883–897.

EFSA (2009). Guidance of the scientific committee on use of the benchmark dose approach in risk assessment. *EFSA Journal*, 7(6):1150.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

Gondane, A. and Itkonen, H. M. (2023). Revealing the history and mystery of rna-seq. *Current Issues in Molecular Biology*, 45(3):1860–1874.

Gordon, K. (2001). The oecd guidelines and other corporate responsibility instruments: a comparison.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., and Maire, P. (2008). The hill equation: a review of its capabilities in pharmacological modelling. *Fundamental & clinical pharmacology*, 22(6):633–648.

Gu, X., Albrecht, W., Edlund, K., Kappenberg, F., Rahnenführer, J., Leist, M., Moritz, W., Godoy, P., Cadenas, C., Marchan, R., et al. (2018). Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes. *Archives of toxicology*, 92:3505–3515.

Haber, F. (1924). Zur geschichte des gaskrieges. *Fuenf vortraege aus den jahren 1920–1923*, pages 76–92.

Hill, A. V. (1910). The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *The Journal of Physiology*, 40:iv–vii.

Hothorn, L. A. (2016). *Statistics in toxicology using R*. CRC Press.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Kappenberg, F., Duda, J. C., Schürmeyer, L., Gül, O., Brecklinghaus, T., Hengstler, J. G., Schorning, K., and Rahnenführer, J. (2023). Guidance for statistical design and analysis of toxicological dose–response experiments, based on a comprehensive literature review. *Archives of Toxicology*, 97(10):2741–2761.

Kappenberg, F. and Rahnenführer, J. (2023). Information sharing in high-dimensional gene expression data for improved parameter estimation in concentration-response modelling. *Plos one*, 18(10):e0293180.

Krug, A. K., Kolde, R., Gaspar, J. A., Rempel, E., Balmer, N. V., Meganathan, K., Vojnits, K., Baquié, M., Waldmann, T., Ensenat-Waser, R., et al. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Archives of toxicology*, 87:123–143.

Kurzrock, R., Lin, C.-C., Wu, T.-C., Hobbs, B. P., Pestana, R. C., and Hong, D. S. (2021). Moving beyond 3+ 3: the future of clinical trial design. *American Society of Clinical Oncology Educational Book*, 41:e133–e144.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: a likelihood approach. *Biometrics*, pages 673–684.

Patro, R., Duggal, G., and Kingsford, C. (2015). Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment. *BioRxiv*, 10:021592.

Phillips, J. R., Svoboda, D. L., Tandon, A., Patel, S., Sedykh, A., Mav, D., Kuo, B., Yauk, C. L., Yang, L., Thomas, R. S., et al. (2019). Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782.

Pinheiro, J., Bornkamp, B., Glimm, E., and Bretz, F. (2014). Model-based dose finding under model uncertainty using general parametric models. *Statistics in medicine*, 33(10):1646–1661.

Pinheiro, J. C., Bretz, F., and Branson, M. (2006). Analysis of dose–response studies—modeling approaches. In *Dose finding in drug development*, pages 146–171. Springer.

Ritz, C. (2010). Toward a unified approach to dose–response modeling in ecotoxicology. *Environmental Toxicology and Chemistry*, 29(1):220–229.

Ritz, C., Baty, F., Streibig, J. C., and Gerhard, D. (2015). Dose-response analysis using r. *PloS one*, 10(12):e0146021.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140.

Rogatko, A., Babb, J. S., Tighiouart, M., Khuri, F. R., and Hudes, G. (2005). New paradigm in dose-finding trials: patient-specific dosing and beyond phase i. *Clinical cancer research*, 11(15):5342–5346.

Rogatko, A., Schoeneck, D., Jonas, W., Tighiouart, M., Khuri, F. R., and Porter, A. (2007). Translation of innovative designs into phase i trials. *Journal of Clinical Oncology*, 25(31):4982–4986.

Sánchez, A. and de Villa, M. (2008). A tutorial review of microarray data analysis. *Bioinformatics*, 1(1):1–55.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441–448.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., et al. (2023). Next-generation sequencing technology: Current trends and advancements. *Biology*, 12(7):997.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

Seber, G. A. and Wild, C. J. (2003). Nonlinear regression. hoboken. *New Jersey: John Wiley & Sons*, 62(63):1238.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2020). Functional horseshoe priors for subspace shrinkage. *Journal of the American Statistical Association*, 115(532):1784–1797.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, pages 925–937.

Su, H., Haque, M., Becker, S., Edlund, K., Duda, J., Wang, Q., Reißing, J., Marschall, H.-U., Candels, L. S., Mohamed, M., et al. (2023). Long-term hypercaloric diet exacerbates metabolic liver disease in pnpla3 i148m animals. *Liver International*.

Tamhane, A. C., Hochberg, Y., and Dunnett, C. W. (1996). Multiple test procedures for dose finding. *Biometrics*, pages 21–37.

Ting, N., Chen, D.-G., Ho, S., Cappelleri, J. C., et al. (2017). *Phase II clinical development of new drugs*. Springer.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wetterstrand, K. A. (Last Updated: November 1 2021). The cost of sequencing a human genome. `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`. Accessed: January 5 2024.

Wheeler, M. W., Lim, S., House, J. S., Shockley, K. R., Bailer, A. J., Fostel, J., Yang, L., Talley, D., Raghuraman, A., Gift, J. S., et al. (2023). Toxicr: A computational platform in r for computational toxicology and dose–response analyses. *Computational Toxicology*, 25:100259.

Wiemann, P. and Kneib, T. (2021). Adaptive shrinkage of smooth functional effects towards a predefined functional subspace. *arXiv preprint arXiv:2101.05630*.

Part II

Publications

Article 1

Received: 19 August 2020 | Revised: 11 November 2021 | Accepted: 12 January 2022

DOI: 10.1002/bimj.202000250

Biometrical Journal

RESEARCH ARTICLE

Model selection characteristics when using MCP-Mod for dose–response gene expression data

Julia C. Duda | **Franziska Kappenberg** | **Jörg Rahnenführer**

Department of Statistics, TU Dortmund University, Dortmund, Germany

Correspondence
Julia C. Duda, Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany.
Email: duda@statistik.tu-dortmund.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: LivSysTransfer (031L0119); Deutsche Forschungsgemeinschaft, Grant/Award Number: RTG 2624 (427806116)

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

We extend the scope of application for MCP-Mod (Multiple Comparison Procedure and Modeling) to in vitro gene expression data and assess its characteristics regarding model selection for concentration gene expression curves. Precisely, we apply MCP-Mod on single genes of a high-dimensional gene expression data set, where human embryonic stem cells were exposed to eight concentration levels of the compound valproic acid (VPA). As candidate models we consider the sigmoid $E_{\max}$ (four-parameter log-logistic), linear, quadratic, $E_{\max}$, exponential, and beta model. Through simulations we investigate the impact of omitting one or more models from the candidate model set to uncover possibly superfluous models and to evaluate the precision and recall rates of selected models. Each model is selected according to Akaike information criterion (AIC) for a considerable number of genes. For less noisy cases the popular sigmoid $E_{\max}$ model is frequently selected. For more noisy data, often simpler models like the linear model are selected, but mostly without relevant performance advantage compared to the second best model. Also, the commonly used standard $E_{\max}$ model has an unexpected low performance.

KEYWORDS
dose–response curves, gene expression, MCP-mod, model selection, toxicology

1 | INTRODUCTION

In drug development, two major steps are of interest when a new compound is examined. First, changes in the dose or concentration of the compound are intended to cause changes in the response. Once this relation is established, the precise modeling of the dose–response curve is the next goal. It aims at finding the target dose for the confirmatory Phase III trials.

If multiple comparison procedures (MCPs) are used for signal detection, this can lead to less flexibility as target dose estimation is restricted to the tested dose levels. One major methodological advancement in this field is the Multiple Comparison Procedure and Modeling (MCP-Mod) approach by Bretz et al. (2005). It combines MCP and a modeling (Mod) step by proposing a multistage procedure. MCP-Mod received a positive qualification opinion and a "fit for purpose" designation by the EMA and FDA in 2014 and 2016, respectively, as statistical methodology to analyze Phase II dose-finding studies under model uncertainty (European Medicines Agency, 2015; Food and Drug Administration, 2016).

This work extends the usual scope of application of MCP-Mod from clinical Phase II to gene expression data. As a practical example, human embryonic stem cells are analyzed (O'Quigley et al., 2017, Chap. 12.3). Valproic acid (VPA) is used for treating epilepsy but it is known to be embryo-toxic when taken in the first trimenon of pregnancy (Genton et al., 2006). The MCP-Mod framework can help to gain insights on concentration–response relationships between the concentration of VPA and gene activity.

Specifically, we are interested in two aspects of MCP-Mod when applied on concentration–response data: the detection of genes where VPA has an effect on the dose–response curve (power) and model selection. We investigate these properties in analyses on real and on simulated data. Further, the model performance or goodness-of-fit of selected models is evaluated to identify which models are suitable for gene expression dose–response data.

Model selection and model performance differ substantially in the underlying theory. In model selection a statistical model from a set of candidate models has to be selected, given a data set. The aim is to select the model that represents the true, unknown model function best (Chap. 1 of Claeskens & Hjort, 2008; Schorning et al., 2016)). In addition to selecting the best model among the candidates, we also aim at identifying necessary or dispensable models. Therefore, we use the goodness-of-fit measure R^2_{adj}. We combine the three aspects power, model selection, and goodness-of-fit in a newly proposed score that summarizes the suitability of a model set. This approach is applied on the VPA data set and on simulated data.

In the context of clinical Phase II trials, model uncertainty for dose–response modeling is considered to increase precision in target dose estimation—Ting (2006), Wheeler and Bailer (2009), Bornkamp et al. (2011) among many others. In Phase II trials, decisions on the model set can be based on expert knowledge and concentrate on a single compound and dose–response relationship. For gene expression data, model selection must be considered for thousands of genes simultaneously and it is not straightforward to find or use prior knowledge on the dose–response profile of each gene. House et al. (2017) and Filer et al. (2016) propose experimental pipelines that include concentration–response modeling and model selection for toxicological gene expression data. They consider a flat model, the sigmoid $E_{\max}$ model with all four parameters or with the lower asymptote fixed to zero, and a gain–loss model that is similar to the beta model considered here. However, detailed investigations on the necessity of model selection and on appropriateness of candidate model sets for gene expression concentration–response data are lacking, which motivates our work.

This paper is structured as follows. The VPA data set is introduced in Section 2. The statistical methodology including MCP-Mod and both established performance measures and a newly proposed one are presented in Section 3. Our analysis procedures and results that are based on the VPA data set are presented in Section 4. Different controlled simulation setups and corresponding results follow in Section 5. Final conclusions are summarized in Section 6. Source code to reproduce the results is available as Supporting Information on the journals web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo).

2 | GENE EXPRESSION DATA SET

The data set was first presented in the study of Krug et al. (2013), where VPA is applied, among others, to human embryonic stem cells (hESC). VPA is widely used to treat different forms of epilepsy. However, it is linked to an increased incidence in congenital abnormalities (Cotariu & Zaidman, 1991). Krug et al. (2013) state that identifying changes in the transcriptome induced by toxic substances illustrates interesting mechanistic insights.

Gene expression levels of the hESCs are measured repeatedly for different concentrations, using the GeneChip R Human Genome U133 Plus 2.0. The data are preprocessed with the Robust Multi-Chip Average algorithm by Irizarry et al. (2003), such that the expression data are on the logarithmic scale with base 2.

The data set contains $G = 54{,}675$ probe sets, which will be referred to as genes in the following, for simplicity. For every gene, expression values corresponding to the concentrations $d_1 = 0$, $d_2 = 25$, $d_3 = 150$, $d_4 = 350$, $d_5 = 450$, $d_6 = 550$, $d_7 = 800$, and $d_8 = 1000$ μM VPA are available. For the control level d_1, $n_1 = 6$ replicates were measured. For all other concentrations there are $n_2 = \cdots = n_8 = 3$ replicates. There are $N = 27$ measurements per gene. The replicates are biological replicates since different cells were used for each experiment. Due to functional relationships between genes, we cannot assume independence between the measurements from different genes. Further, six or three replicates per concentration is small for statistical modeling approaches. These problems are addressed in Section 4.

3 | MCP-MOD METHODOLOGY AND PERFORMANCE MEASURES

In this section, the methodology is presented. First, the MCP-Mod approach is outlined. Then, the performance measures precision and recall for evaluating the model selection in MCP-Mod are explained. Additionally, the newly proposed measure $S_{\mathcal{M}}$ is presented.

TABLE 1 Dose–response models $f(d,\theta)$, their standardized versions $f^0(d,\theta^0)$, and the guesstimates for θ^0 for the analysis. For the beta model B is defined as $B(\delta_1,\delta_2)=(\delta_1+\delta_2)^{\delta_1+\delta_2}/(\delta_1^{\delta_1}\delta_2^{\delta_2})$ and $D=1200$

Model	$f(d,\theta)$	$f^0(d,\theta^0)$	θ^0
$E_{\max}$	$E_0+E_{\max}d/(ED_{50}+d)$	$d/(ED_{50}+d)$	$ED_{50}\in\{100\}$
Sigmoid $E_{\max}$	$E_0+E_{\max}d^h/(ED_{50}^h+d^h)$	$d^h/(ED_{50}^h+d^h)$	$ED_{50}=450$, $h=5.117$
Exponential	$E_0+E_1\{\exp(d/\delta)-1\}$	$\exp(d/\delta)-1$	$\delta\in\{144.455\}$
Linear	$E_0+\delta d$	d	$\emptyset$
Quadratic	$E_0+\beta_1 d+\beta_2 d^2$	$d+(\beta_2/\lvert\beta_1\rvert)d^2$	$\delta=\beta_2/\lvert\beta_1\rvert=-0.001$
Beta	$E_0+E_{\max}B(\delta_1,\delta_2)(d/D)^{\delta_1}\cdot(1-d/D)^{\delta_2}$	$(d/D)^{\delta_1}(1-d/D)^{\delta_2}$	$\delta_1=2$, $\delta_2=1$

3.1 | MCP-Mod

The MCP-Mod approach was originally developed by Bretz et al. (2005) to model dose–response relationships in Phase II clinical trials under model uncertainty. For details see also Xun and Bretz (2017) and Bornkamp et al. (2009).

The MCP-Mod methodology comprises two analysis steps. First, in the MCP-step, a statistically significant signal in a gene is determined by an optimal-contrast test for a prespecified set of candidate models. If such a signal is found for at least one model, a significant result of the multiple comparison procedure (signifMCP) is present for the gene. This means that an effect of VPA on the gene activity is established. The second step, Mod, refers to the modeling. From the set of models, for which a signifMCP has been established, one model fit is chosen and used as final fit for the data. Alternatively, model averaging can be performed.

Denote d_1 as placebo concentration and $d_2<\ldots<d_k$ as active concentrations with n_i replicates. For concentration $i=1,\ldots,k$ and $j=1,\ldots,n_i$, $N=n_1+\cdots+n_k$, the (preprocessed) expression values are modeled as

$$y_{ij}=\mu(d_i)+\varepsilon_{ij},\ \varepsilon_{ij}\overset{i.i.d.}{\sim}\mathcal{N}(0,\sigma^2), \tag{1}$$

with homogeneous variance $\sigma^2>0$. The mean response $\mathrm{E}(y_{ij})=\mu_i=f(d_i,\theta)$ at concentration d_i is assumed to follow a concentration–response model with parameter vector θ and ε_{ij} as independent errors.

For the MCP step, a set $\mathcal{M}$ of M candidate models needs to be prespecified. Models commonly used for dose–response relationships are summarized in Table 1.

All models summarized in the first column of Table 1 can be reformulated as

$$f(d,\theta)=\theta_0+\theta_1 f^0(d,\theta^0), \tag{2}$$

(see second column of Table 1), where $f^0(d,\theta^0)$ is the standardized version of a model. Introduction of the standardized model shape concept is crucial for choosing optimal contrasts in the MCP step, as their choice is scale invariant.

It remains to determine initial guesses for the parameter θ^0. In practice, for a Phase II study, careful considerations and prior knowledge on expected percentages of maximal effects at certain doses are translated into guesstimates for θ^0. Here, the large number of genes makes individual, gene-dependent decisions on θ^0 difficult. We therefore use the same guesstimates for all genes. Figure 1 displays the (rescaled) model shapes $f^0(d,\theta^0)$ used for the analysis. The guesstimates are listed in Table 1.

To the best of our knowledge there is little preliminary work on dose–response model selection in the context of gene expression data (Filer et al., 2016; House et al., 2017). In toxicology, often monotone dose–response relationships are assumed. Especially the $E_{\max}$ model, a special case of the sigmoid $E_{\max}$ model with $h=1$, was found to be appropriate for the majority of dose–response relationships in a large meta-analysis of clinical dose–response studies (Thomas et al., 2014). The inclusion of these two monotonic models in the candidate model set is therefore obligatory. The linear model is added as a reference or baseline model. For genes where the true underlying model might be a sigmoid $E_{\max}$ model, but at the maximal considered dose, the turning point has not yet been reached, the exponential model might be more suitable. The quadratic and the beta model are included as nonmonotone shapes. They are similar to the gain–loss model used by Filer et al. (2016). There might be a nonmonotone relationship between concentration and gene activity, for example, for metabolic genes. Such genes might be activated at lower VPA concentrations but successively deactivated at increasing, highly toxic concentrations.

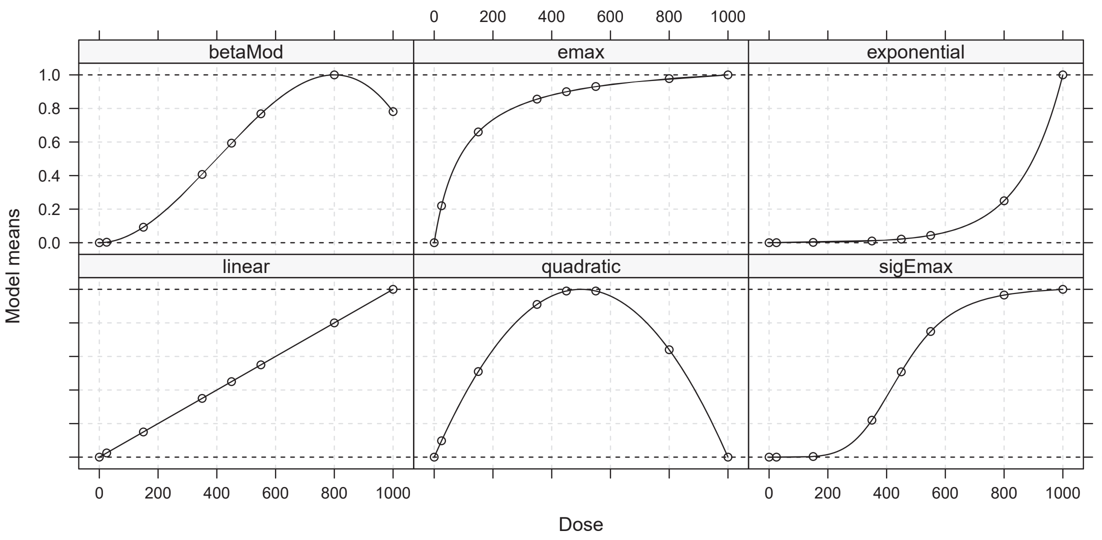


FIGURE 1 Candidate model shapes used for the six concentration–response models

The specific guesstimates for $\boldsymbol{\theta}^0$ for each model are chosen such that a wide range of dose–response shapes is covered. Further, we consider that during the experimental design stage, the concentrations were chosen with the expectation that the dose with the half maximal effect (ED_{50}) is close to 450 and a plateau is reached at concentration 1000, which translates into the second guess that 95% of the maximal effect is reached at concentration 800. With these two assumptions ($ED_{50} \approx 450$ and $ED_{95} \approx 800$), the guesstimate $\boldsymbol{\theta}^0$ for the sigmoid $E_{\max}$ model can be calculated analytically. For the $E_{\max}$ model, a guess of an ED_{50} of 300 is used. And for the exponential model, an ED_{50} of 700 is assumed.

Each candidate shape, $m = 1, \ldots, M$, defines a respective mean response vector $\boldsymbol{\mu}_m = (\mu_{m1}, \ldots, \mu_{mk})$. For the MCP-step, a contrast t-test as first described by Abelson et al. (1963) is calculated. The test is constructed based on the linear contrast $\boldsymbol{c}_m^\top \boldsymbol{\mu}_m$ where $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mk})^\top$ is chosen to maximize the power of the test for the assumed mean response $\boldsymbol{\mu}_m$ (Bornkamp et al., 2009). This yields the hypotheses $\mathrm{H}_0^m : \boldsymbol{c}_m^\top \boldsymbol{\mu}_m = 0$ and $\mathrm{H}_1^m : \boldsymbol{c}_m^\top \boldsymbol{\mu}_m \neq 0$.

The test statistics for the contrasts are given by

$$T_m = \frac{\sum_{i=1}^{k} c_{mi} \bar{y}_i}{S\sqrt{\sum_{i=1}^{k} c_{mi}^2 / n_i}}, \; m = 1, \ldots, M, \tag{3}$$

where $\bar{y}_i$ is the observed mean at dose d_i and $S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (N - k)$ is the mean squared error. Under H_0 and (1), $(T_1, \ldots, T_m)^\top$ follows a central, multivariate t-distribution.

A dose–response signal is established if $T_{\max} = \max(T_1, \ldots, T_m) > q_{1-\alpha}$, where q_α is the equicoordinate α-quantile of the null distribution. This approach leads to multiple testing adjustment for $\{\mathrm{H}_0^m, \mathrm{H}_1^m\}$ with a strict control of the family wise error rate at level α. The models with $T_m > q_{1-\alpha}$ form the set $\mathcal{M}^* = \{M_1, \ldots, M_L\}$ of L significant models with established signifMCP. The modeling step is only executed if $\mathcal{M}^* \neq \emptyset$, that is, a dose–response signal is established for at least one model.

During the Mod-step, either one fitted model of those that passed the MCP-step can be chosen for a final fit or all fitted models that passed the MCP-step can be averaged. If a single model is selected, criteria as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) as well as the largest test statistic (maxT) can be used to pick a model. For model averaging, standardized weights based on the AIC or BIC can be calculated for the models in $\mathcal{M}$, and the final model is the resulting weighted average of each of the fitted models.

Calculations are done with the `DoseFinding` R package, version 0.9-17, and the statistical software `R`, version 4.0.2 (R Core Team, 2020). For the numerical estimation of the nonlinear parameters, we use the default boundary setting of the `DoseFinding` package. As the maximum concentration is 1000, this leads to boundaries for the ED_{50} parameter of [1, 1500] and [1/2, 10] for the h parameter of the sigmoid $E_{\max}$ model.

3.2 | Measures

In this section, we briefly present for our context the definitions of the evaluation measures precision, recall, and R^2_{adj}. Further, a new measure $\mathcal{S}_{\mathcal{M}}$ is proposed specifically for the context of using MCP-Mod with a fixed candidate set $\mathcal{M}$ on many dose–response data sets.

For a set of genes and a specific model $M \in \mathcal{M}^* = \{M_1, \dots, M_L\}$, the precision is defined as the conditional probability that a model is correct, given that it has been selected. Accordingly, the recall is defined as the conditional probability that a model is selected, given that it is correct (Buckland & Gey, 1994). Formally, we denote

$$\text{precision} = \hat{P}(\text{Model is correct} \mid \text{Model is selected}) = \frac{tp}{tp + fp},$$

$$\text{recall} = \hat{P}(\text{Model is selected} \mid \text{Model is correct}) = \frac{tp}{tp + fn},$$

where tp, fp, and fn are the number of true positives, false positives, and false negatives. Precision and recall values are in the interval [0,1], and a larger value corresponds to a better performance. They can only be evaluated in simulations where the correct model is known.

For a model fit $f(\cdot, \hat{\theta})$ and data y_{ij} of a specific gene, $i = 1, \dots, k$, $j = 1, \dots, n_i$, we use R^2_{adj} defined as

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p}, \quad R^2 = 1 - \frac{SSE}{SST}, \tag{4}$$

where $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - f(d_i, \hat{\theta}))^2$ and $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ is the sum of squared errors and total sum of squares, respectively. The number of parameters is p and the total number of measurements is $N = \sum_{i=1}^{k} n_i$.

We further propose a new measure, the suitability of model set score $\mathcal{S}_{\mathcal{M}}$. It can be used in a descriptive manner when MCP-Mod is applied to dose–response or concentration–response data of many genes. The score balances two desired properties. First, the number of detected signals (signifMCPs) is desired to be large. Additionally, the detected signals are also desired to be clear, that is, to have a large R^2_{adj} value. The score balances the number of detected signals and the model performance, that is, power and goodness-of-fit. It is defined as

$$\mathcal{S}_{\mathcal{M}} = \frac{1}{G} \sum_{g=1}^{G} \mathbb{1}\{\text{Gene } g \text{ has significant MCP after adjustment}\} \cdot R^2_{\text{adj}}, \tag{5}$$

where G denotes the total number of genes and $\mathcal{M}$ the considered set of candidate models. For a given set $\mathcal{M}$, the score summarizes the proportion of genes with significant MCP after adjustment, weighted by the goodness-of-fit of the respective genes. Adjustment means that the false discovery rate (FDR) is controlled using the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995).

In the context of MCP-Mod, this means that for each gene, the smallest p-value from the MCP tests of all candidate models is chosen. Consequently, each gene is represented by a single p-value. These p-values are adjusted with the BH procedure. If a BH-adjusted p-value is below 0.05 then this results in a multiplicity-adjusted significant concentration–response signal. As performance measure, the value of R^2_{adj} of the winner model w.r.t. AIC for the corresponding gene is used. In general, when comparing two values $\mathcal{S}_{\mathcal{M}_1}$ and $\mathcal{S}_{\mathcal{M}_2}$, the larger value indicates a favorable choice of the candidate set, since both the number of detected signals and the model performance are taken into account. For improved clarity, in addition both the proportion of genes with detected signal and the mean R^2_{adj} of the fit of the winner models corresponding to those genes will be reported.

4 | DATA-BASED ANALYSIS

In this section, setups and results of the data-based analyses are presented. In the following, they are referred to as Analysis I and Analysis II. In Analysis I, MCP-Mod is applied on the real VPA data set and an overview on model selection results

TABLE 2 Overview of the analyses scenarios and their respective data generation details (Section 4) as well as for the simulation study of Section 5

Part		Data (generation)
Analysis I	main	Original VPA data
Analysis II	LOMO	Original VPA data
Simulation		$n_1 = \cdots = n_8 \in \{3, 5, 10\}$
		$\sigma = q(0.5) \cdot \text{range}$ ($\sigma = q_{\text{null}}(0.5)$ for null-model)

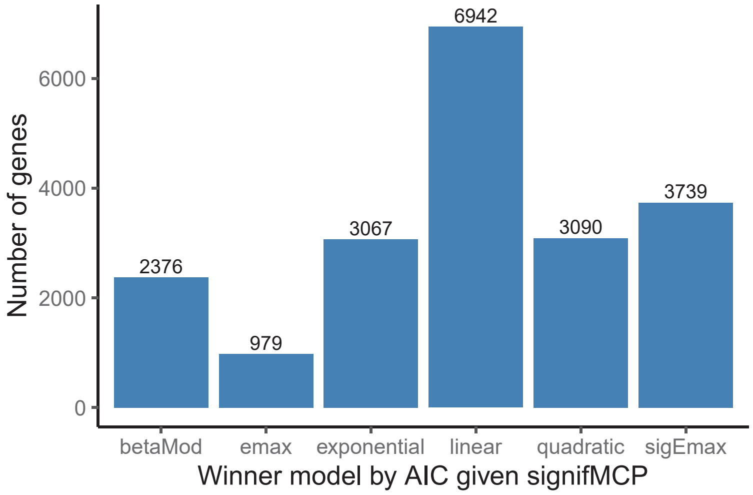


FIGURE 2 Distribution of winner counts per model

is provided. An additional goal is to check if any model can be omitted from the candidate model set because it can be easily substituted by another model. Analysis I is extended by Analysis II through leave-one-model-out (LOMO) analyses. These include that the entire analysis of Analysis I is repeated several times, and each time one of the candidate models of $\mathcal{M}$ is omitted. For an overview of the different analyses and the simulation, see Table 2.

4.1 | Setup for Analysis I

In Analysis I, MCP-Mod is applied independently on each gene of the VPA data set. As we cannot assume only increasing or decreasing effects, each gene is tested with two-sided contrast tests with significance level $\alpha = 0.05$.

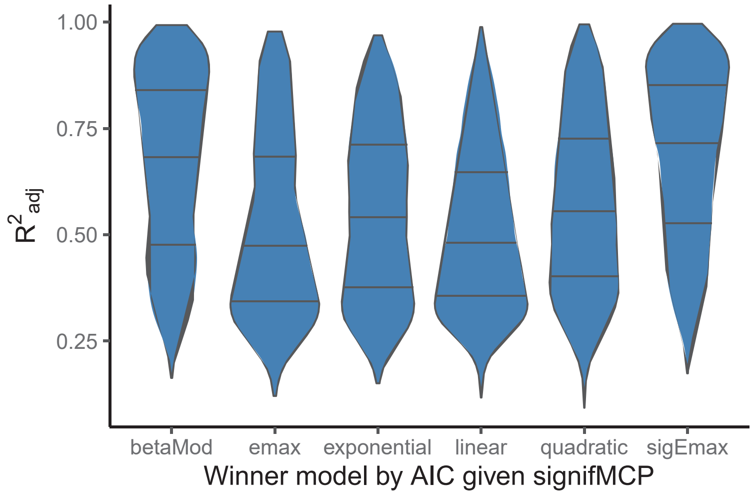
We apply multiplicity adjustment between genes by controlling the FDR using the BH procedure as described in Section 3.2. For each gene, if a dose–response signal is detected and hence at least one model passes the MCP-step, the AIC is used as model selection criterion. For small sample sizes, Schorning et al. (2016) show that the AIC outperforms the BIC, especially if the true underlying model is a complex one among the considered models. There are $N = 27$ observations per gene in the VPA data set. Thus we use the AIC to avoid too low selection counts of possibly correct, more complex models. In our analysis we will see that even with the AIC the simple linear model is often selected.

4.2 | Results for Analysis I

Of the 54,675 genes, when controlling the FDR, 20,193 (36.9%) genes pass the MCP-step, that is, their concentration–response profile significantly differs from a flat profile. VPA has a significant effect on the activation (deactivation) of these genes. For each gene one winner model is selected by AIC as a final fit (Figure 2). The linear model is selected most often (34.4%), because the AIC penalizes more complex models. However, Figure 3 clearly shows that the linear model performs comparatively poorly with respect to the R^2_{adj}.

The popular $E_{\max}$ model (cf. Thomas et al., 2014, among many others) wins the fewest times and when it does win, its fit has low R^2_{adj} values (Figures 2, 3). Figure 4 shows the distribution of model winner counts w.r.t. AIC stratified by R^2_{adj}. Less noisy genes are represented by the rightmost plot. Assuming (1), we refer to more (less) noisy genes as genes whose

FIGURE 3 Performance of winner models



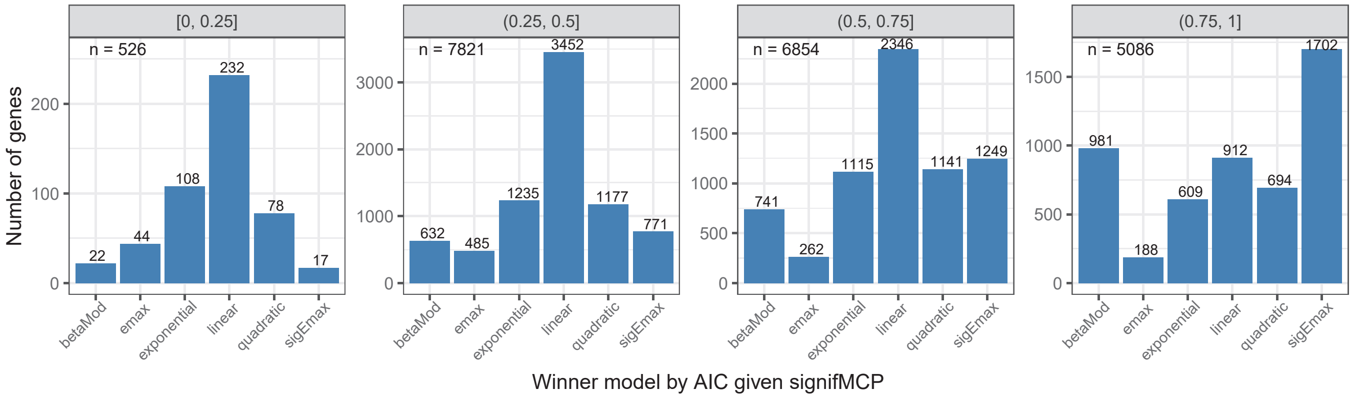


FIGURE 4 Distribution of winner counts per model stratified by R^2_{adj}

underlying model has larger (smaller) standard deviations σ in relation to the response range, which leads to smaller (larger) R^2_{adj} values. While the sigmoid E_{max} and the beta model win most often for the least noisy stratum, the E_{max} model is chosen rarely, regardless of the strata. The nonmonotone beta and quadratic model are chosen considerably often. For more noisy genes the linear model is preferred. For these genes, none of the models explains a lot of variance, which favors the linear model in terms of AIC. Hence, if the linear model is selected by AIC, one should hesitate to assume a true linear concentration–response relationship. Some example fits are visualized in Figure 5.

To ensure that the low number of E_{max} winners is not only due to too strict parameter constraints in the numerical optimization, we visualize the ED_{50} parameter estimates for genes where the E_{max} model won (Appendix, Figure A1). There is no evidence that the poor performance of the E_{max} model is due to optimization constraints, but instead due to the often low ED_{50} estimates. For an E_{max} model, a low ED_{50} translates to an early plateau, which can lead to an SSE close to the SST and therefore to a small R^2_{adj}.

It is also of interest if any of the models in the candidate set is redundant such that it can be substituted by another model. Removing such a model from the candidate set would increase power as the number of hypotheses would be decreased in the MCP-step of each gene. If for many genes the R^2_{adj} for the winner model and the second best model differ substantially, the winning model should be considered for future analyses. If the quadratic model is the winning model with a good fit, many genes cannot be modeled well by the second best model (Figure 6).

The sigmoid E_{max} and the beta model performances also differ by a considerable amount to the second best model's performance across the whole range of explained variance. The E_{max} and the exponential model can mostly be replaced by other models without substantial loss in R^2_{adj}. This especially applies to genes with larger explained variance. The linear model can always be replaced with minimal loss in explained variance, as it is a special case of the E_{max} model and the quadratic model.

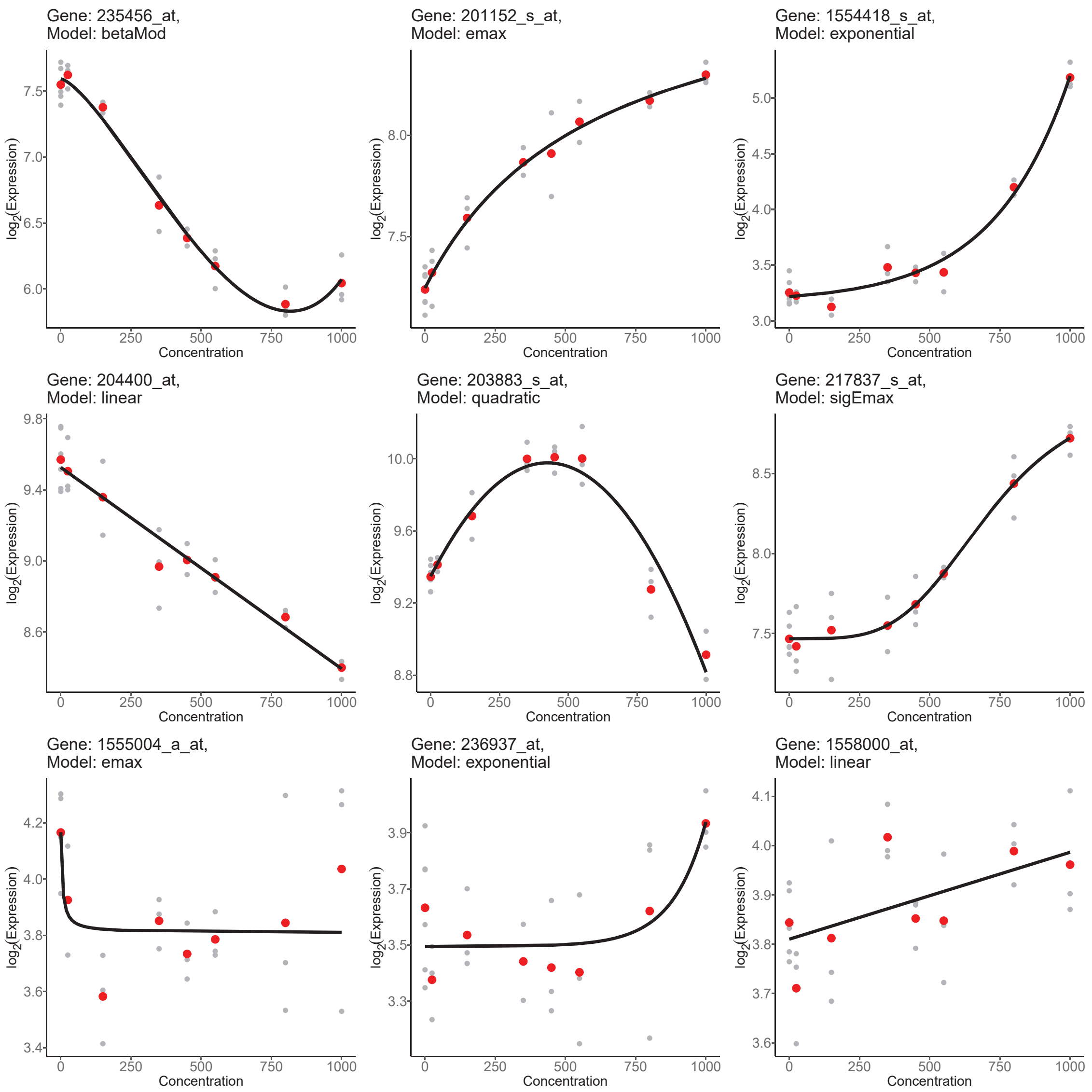


FIGURE 5 Example selection of nonnoisy (rows 1 and 2) and noisy (row 3) genes of the VPA data set with significant concentration–response model, with added fit of the model selected as winner w.r.t. AIC. Gray dots represent single response values, and the red dots indicate the mean responses per concentration. Each model has an R^2_{adj} of at least 0.75 (rows 1 and 2) or below 0.25 (row 3)

4.3 | Setup for Analysis II

Analysis II offers further insights into possibly expendable models in the candidate set. The analysis setup is similar to the one of Analysis I. Analysis I is redone several times, but each time one model from the candidate model set is omitted. We refer to these as LOMO analyses.

4.4 | Results for Analysis II

The number of FDR adjusted significant concentration–response relationships is similar to the main analysis where no model is left out (Table 3). This finding is consistent with the results of Pinheiro et al. (2006). If the quadratic model is omitted from the candidate model set, the total number of signifMCPs increases at the cost of reduced mean R^2_{adj} for the remaining genes. This is due to the different, rarely appropriate shape of the quadratic model compared to all other models. Measured by the $S_{\mathcal{M}}$ score, it is proposed to drop the $E_{\max}$ model from the candidate model set (indicated in bold). The

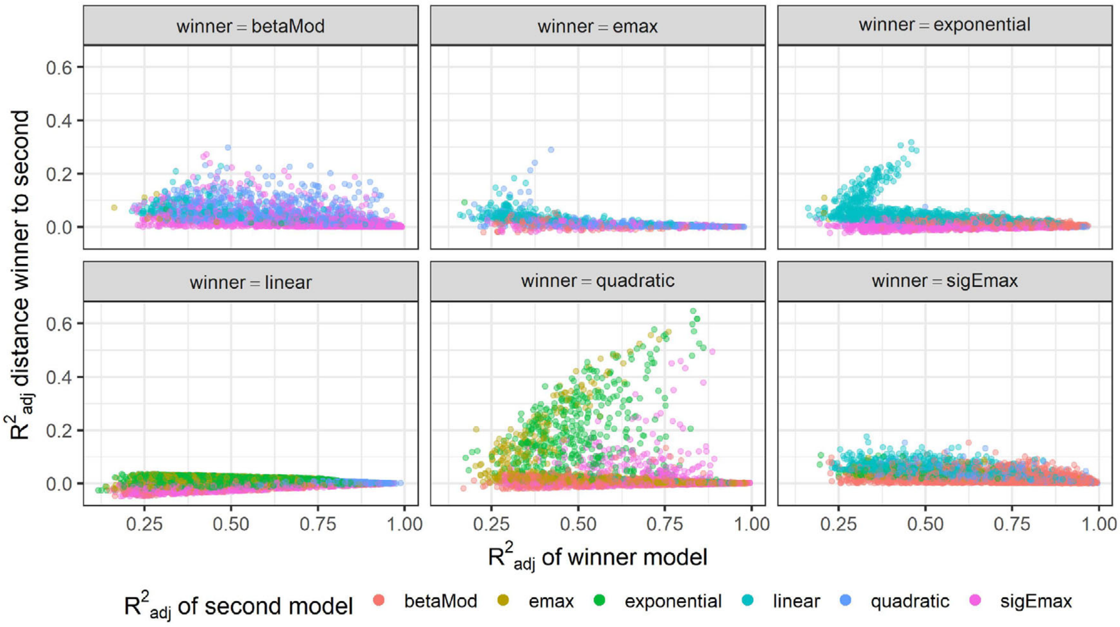


FIGURE 6 Scatter plots stratified by winner model (by AIC) showing the R^2_{adj} value of the winner on the x-axis and the difference to the R^2_{adj} value of the second best model (by AIC) on the y-axis

TABLE 3 Total number of FDR adjusted significant genes in each LOMO analysis, number of gained and lost genes compared to the case when no model is left out (Analysis I), $\mathcal{S}_{\mathcal{M}}$ score, rate of FDR adjusted significant genes, and mean R^2_{adj} among significant genes. Largest $\mathcal{S}_{\mathcal{M}}$ score indicated in bold

Model	Total	Gained	Lost	$\mathcal{S}_{\mathcal{M}}$	signifMCP rate	mean R^2_{adj}
Sigmoid $E_{\max}$	20,122	61	132	0.2114	0.3680	0.5743
Quadratic	20,459	697	431	0.2119	0.3742	0.5664
Beta	20,221	150	122	0.2120	0.3698	0.5732
Exponential	20,164	529	558	0.2120	0.3688	0.5748
Linear	20,178	91	106	0.2127	0.3691	0.5763
None	20,193	0	0	0.2134	0.3693	0.5778
$E_{\max}$	20,349	330	174	**0.2138**	0.3722	0.5745

sigmoid $E_{\max}$ model, which contains the $E_{\max}$ model as a special case, decreases the score the most when removed from the candidate set.

We are further interested by which model an originally selected model after its omission is typically substituted in the modeling step (Figure 7). The beta model is selected more often, if the sigmoid $E_{\max}$ model is removed and vice versa. If the often selected linear model is omitted, the exponential model is most often replacing it.

Two additional evaluations regarding the validity of the $\mathcal{S}_{\mathcal{M}}$ score were conducted. First, a copy of the VPA data set was generated and all 3067 genes where the exponential model won by AIC were removed and the LOMO analyses were repeated. As expected, in this artificial scenario the $\mathcal{S}_{\mathcal{M}}$ score suggests to drop the exponential model (Table 4, second column from the left).

Second, Analysis I was repeated but with a single model as candidate model (Table 4). When a single candidate model is used, the $\mathcal{S}_{\mathcal{M}}$ score is always smaller than when only one or no model is omitted from the candidate model set and the original VPA data are used. The lowest score of 0.0825 is obtained if the quadratic model is the only candidate model. This is because the quadratic model shape passes the MCP-step for only 12.27% of the genes. When using only one candidate model, the sigmoid $E_{\max}$ model has the largest score of 0.2036.

The absolute differences in $\mathcal{S}_{\mathcal{M}}$ scores might appear small but must not be misinterpreted as irrelevant. For the artificial scenario where genes with the exponential model as winner model are removed, omitting the exponential model from the candidate model set is considered reasonable by construction. Therefore, the corresponding difference in the $\mathcal{S}_{\mathcal{M}}$ score of

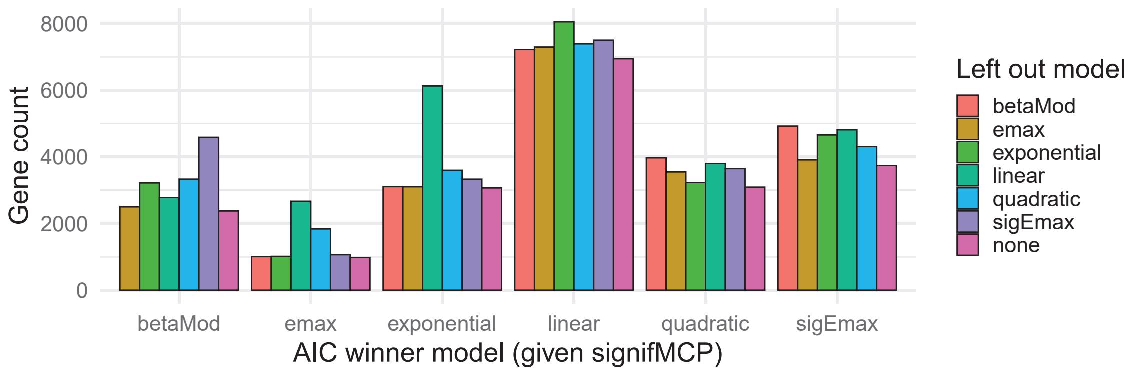


FIGURE 7 Absolute count of selections by AIC in the modeling step of each model that had a signifMCP in the MCP-step for each LOMO scenario

TABLE 4 $\mathcal{S}_\mathcal{M}$ score, rate of significant genes, and mean R^2_{adj} among significant genes for two added analyses: The LOMO analyses on the modified VPA data set where genes with exponential winner model are removed and Analysis I repeated but with only one model in the candidate model set. Largest $\mathcal{S}_\mathcal{M}$ score indicated in bold

	LOMO analyses on VPA data set without exponential winner genes			Only model in candidate model set on original VPA data set		
Model	$\mathcal{S}_\mathcal{M}$	**signif. genes (%)**	**mean R^2_{adj}**	$\mathcal{S}_\mathcal{M}$	**signif. genes (%)**	**mean R^2_{adj}**
sig. $E_{\max}$	0.1789	0.3062	0.5844	**0.2036**	0.3771	0.5399
Beta	0.1795	0.3077	0.5834	0.2013	0.3647	0.5520
Quadratic	0.1795	0.3117	0.5758	0.0825	0.1227	0.6722
Linear	0.1803	0.3071	0.5869	0.1855	0.3768	0.4922
None	0.1810	0.3078	0.5882	-	-	-
$E_{\max}$	0.1813	0.3097	0.5854	0.1690	0.3235	0.5225
Exponential	**0.1834**	0.3173	0.5778	0.1669	0.3011	0.5544

0.0024 can be interpreted as relevant. Only using the sigmoid $E_{\max}$ model compared to using the full candidate model set differs by 0.0098, which can hence be viewed as a relevant difference such that it would not suffice to use a single model. The interpretation of the $\mathcal{S}_\mathcal{M}$ score is not straightforward, which is discussed in Section 6.

5 | SIMULATION-BASED ANALYSIS

The simulation gives insights on the effect of the number of replications per concentration level while standard deviation of the noise is fixed (Table 2). Opposed to the data-based analysis, the correct model is known such that precision, recall, and goodness-of-fit can be evaluated.

5.1 | Setup

Concentration–response data sets are generated for 10,000 genes for each of the six considered models and for the null case, as well as for three different numbers of replicates n_i and a fixed standard deviation to range ratio (see Table 2). Details on how the range and standard deviation are chosen are explained in the following. The null case means that a constant model is used to generate the data. In order to have a realistic data generation process, it is based on the real VPA data set. For each considered n_i, a data set structurally similar to the VPA data set but with 70,000 genes, 10,000 for each of the six nonflat models, and 10,000 for the flat null model, is generated as follows.

Consider a model $f = f(d, \theta) \in \mathcal{F}$ where $|\mathcal{F}| = 7$ and for the null case, $f = f(d, \theta) = f(d, c) = c > 0$. The assumed ratio of standard deviation to range denoted by $q(0.5)$ is explained below. Define $\mathcal{G}^*(f)$ as the set of all genes g for which

TABLE 5 Summary of the simulation results, stratified by correct model and chosen model, respectively. Signif. genes (%) is the rate of FDR adjusted, detected dose–response signals among the 10,000 generated dose–response data for each model, with n_i replicates at each dose level. For the recall rate, the model is the correct model. For the precision rate, the model is the selected model

Measure	n_i	Beta	$E_{\max}$	Exponential	Flat	Linear	Quadratic	sig. $E_{\max}$
Signif. genes (%)	3	0.989	0.982	0.979	0.037	0.988	0.988	0.997
	5	1.000	1.000	1.000	0.042	1.000	1.000	1.000
	10	1.000	1.000	1.000	0.045	1.000	1.000	1.000
Recall	3	0.447	0.496	0.561	0.963	0.705	0.543	0.471
	5	0.567	0.552	0.657	0.958	0.750	0.638	0.567
	10	0.712	0.616	0.754	0.955	0.770	0.723	0.717
Precision	3	0.582	0.658	0.682	0.925	0.416	0.547	0.510
	5	0.627	0.746	0.745	0.999	0.529	0.565	0.586
	10	0.676	0.790	0.822	1.000	0.713	0.614	0.689

model f was the selected winner model by AIC in the VPA data set in Analysis I. Further, $\underline{\mathcal{G}}(f)$ is a sample of 10,000 genes drawn with replacement from $\mathcal{G}^*(f)$. For a gene $g \in \underline{\mathcal{G}}(f)$, the true underlying concentration–response relationship is assumed to be the model fit $f^{(g)}(d, \hat{\theta})$ of model f on the VPA data of gene g. For the null model, the mean response $\bar{y}^{(g)}$ is used as true value for c.

Given this true concentration–response relationship of gene g, noise is added to generate a data set according to the model equation (1). For concentration levels $d \in \{d_1, \dots, d_8\}$ used in the original experiment (see Section 2), generate $y_{ij}^{(g)} = f^{(g)}(d_i, \hat{\theta}) + e_{ij}, j = 1, \dots, n_i$. The added noise values e_{ij} are independently drawn from $\varepsilon \sim \mathcal{N}(0, (\sigma(f^{(g)}, s))^2)$. If $f^{(g)}$ is not the null case, then $\sigma(f^{(g)}, s) := \text{range}(f^{(g)}) \cdot q(s)$, where the range for a gene with model f is calculated as $\text{range}(f^{(g)}) := \max\limits_{d\in\{d_1,\dots,d_8\}} (f^{(g)}(d, \hat{\theta})) - \min\limits_{d\in\{d_1,\dots,d_8\}} (f^{(g)}(d, \hat{\theta}))$. The term $q(s)$ is the empirical s-quantile of the ratio $S^{(g)}/\text{range}(f^{(g)})$ across all genes g with a detected signal and their respective fits in Analysis I. Here, $(S^{(g)})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}^{(g)} - \bar{y}_i^{(g)})^2/(N-k)$ is the estimated variance for each gene. Hence, for $s = 0.5$, we obtain $q(0.5) = 0.3222$, which is used for all nonflat models and all genes to calculate $\sigma(f,^{(g)} 0.5)$ (Appendix, Figure A2). If $f^{(g)}$ is the null case, then $\text{range}(f^{(g)}) = 0$. In this case, we use $\sigma(f^{(g)}, s) = q_{\text{null}}(s)$, which is the empirical s-quantile of S calculated for nonsignificant genes g of Analysis I. We obtain $q_{\text{null}}(0.5) = 0.1909$ (Appendix, Figure A3).

Using an adapted standard deviation per gene and model is preferred over using a fixed standard deviation, because it allows for comparability between different models and ranges with respect to goodness-of-fit (Kappenberg et al., 2021). Finally, the generated data set is analyzed as the original VPA data set in Analysis I.

5.2 | Results

Table 5 summarizes the results of the simulation w.r.t. signal detection (power), recall, and precision. For $n_i = 3$, which mimics the conditions in the real VPA data set, a signal is almost always detected if it is present. However, the recall and precision rates for nonlinear and nonflat models for this scenario are below 0.69. If $n_i = 10$, the rates of these model selection errors are still large, even though the sample size $N = 8 \cdot n_i = 80$ is rather large in the context of toxicology. For example, if the sigmoid $E_{\max}$ model is correct, for 31.1% of the generated dose–response data another model is incorrectly selected. Due to the penalty term of the AIC used in model selection, complex correct models as the sigmoid $E_{\max}$ or the beta model have a comparatively large increase in recall rates when n_i is increased. For comparatively noisy scenarios, these models are rarely selected. The opposite holds for the least complex model, the linear model. It has a comparatively very low precision rate (41.6%) and very high recall rate (70.5%) for $n_i = 3$, but not for $n_i = 10$. Precision values naturally have more practical value, as they give insight on how confident one can be with the model selection. The precision rate increases from 92.5% to 99.9% for the flat model if n_i increases from 3 to 5. For nonflat models, the precision rate does not exceed 82.2% at any n_i.

In practice, often one is not mainly interested in selecting the true underlying model but to have a sufficiently good fit. Figure 8 summarizes the relative loss in model fit by considering the log-ratio in root-mean-square error (RMSE) between the winner and the true model, that is, between the actually selected model and the fitted model if the correct

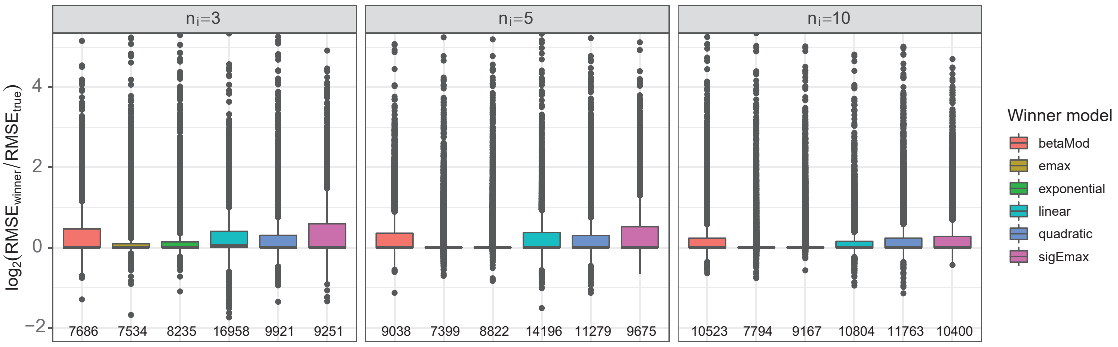


FIGURE 8 Distributions of $\log_2(RMSE_{\text{winner}}/RMSE_{\text{true}})$ in the simulation. $RMSE_{\text{true}}$ is the root mean squared error (RMSE) if the correct model is fitted and $RMSE_{\text{winner}}$ the RMSE of the selected winner model. For 108 genes, the log-ratio is greater than 5 and not displayed

model is selected. For both, the RMSE is calculated with respect to the true dose–response curve that is known in the simulation. If the correct model is actually selected, this ratio is 0, because the true and selected model are the same. If the ratio is greater than one, then the selected model differs from the correct model and has a worse RMSE. The ratio of the RMSE values is independent of n_i and of the range of the respective gene. It only captures the effect of the model selection.

In general, the relative loss in RMSE decreases with increasing n_i but is still present for $n_i = 10$, although $N = 8 \cdot n_i = 80$ is already a large sample size in toxicology. For $n_i = 3$ the median of the log-ratio is 0.0000 for all winner models, but 0.0543 for the linear model. This demonstrates the low precision of the linear model for small n_i. For small $n_i = 3$, the penalty of the AIC is comparatively large, yielding too many selections of the simpler linear model in cases where a more complex model might be required. If a more complex model such as the beta or sigmoid $E_{\max}$ model is selected, the ratio's upper quartile are largest with $2^{0.464} = 1.480$ and $2^{0.594} = 1.509$, respectively. Hence, for 25% of the generated genes where the sigmoid $E_{\max}$ model is selected, the selection is not correct and the RMSE is at least 50.9% larger than the RMSE of the correct model. For the $E_{\max}$ and for the exponential model, the upper quartiles of the ratio are closest to zero for each n_i. With increasing n_i, the penalty term of the AIC becomes comparatively weak. For $n_i = 10$, this heavily affects the linear model. It is selected less often and has log-ratios closely concentrated around zero. For the beta, quadratic, and sigmoid $E_{\max}$ model, the upper quartile of the log-ratios remain comparatively far from 0. If the beta model is selected, for 25% of the generated genes the selection is incorrect and the RMSE is at least $2^{0.235} = 1.177$ times the RMSE if the correct model was selected. Thus, not selecting the correct model results in a noteworthy relative loss in goodness-of-fit, even when larger sample sizes are used in toxicology.

6 | CONCLUSION

In this work, MCP-Mod was used as model selection approach for gene expression concentration–response data. For the data set at hand, human embryonic stem cells were exposed to varying concentrations of VPA. For 54,675 probe sets or genes the expression is measured. The data set indicated that modeling gene expression concentration–response data requires the consideration of several models, that is, a candidate model set. Only considering the popular $E_{\max}$ or sigmoid $E_{\max}$ model might not be sufficient. Especially nonmonotone models like the quadratic model should also be taken into account. When using MCP-Mod, frequent selections of a linear model should not be misinterpreted as evidence for a true, linear concentration–response relationship. A large noise-to-signal ratio, or, more precisely, a large standard deviation to true response range ratio, favors the selection of the linear model. Also, there is typically no notable loss in goodness-of-fit, when instead of the linear model the second best model is used.

Using a newly proposed score, $\mathcal{S}_{\mathcal{M}}$, it was observed that the $E_{\max}$ model can be omitted from the candidate set despite its popularity, as long as the more general sigmoid $E_{\max}$ model is included in the candidate set. Further, the score discourages to omit the linear model, even though it can be easily substituted with respect to goodness-of-fit. Including the linear model in the candidate set aims to detect unclear concentration–response signals rather than modeling detected signals well. If the linear model is omitted, one might fail to identify potentially interesting genes. Simulation studies based on the data set indicate that the confidence in the correctness of the selected model, measured by the precision, is not very high.

Even when increasing the sample size per concentration from 3 to 10, which is very large for this type of toxicological experiments, the precision of nonflat models does not exceed 0.83. Thus, increasing the number of experiments does not increase the precision in model selection proportionally. The relative loss in goodness-of-fit due to model selection mistakes decreases with increasing sample size, but remains notable even for 10 replicates per concentration.

The newly proposed $\mathcal{S}_{\mathcal{M}}$ score served as a help to summarize analysis and simulation results. For a given candidate set, it considers the power, that is, number of detected genes, and the goodness-of-fit of genes with a detected signal simultaneously. Despite its simple form, its interpretability is not straightforward, which allows for improvements. If one does not want to consider both power and goodness-of-fit at the same time but, for example, focuses on optimizing power, the score is not an adequate tool.

The data basis of this work is a single data set, which, despite its size and quality, is an obvious limitation. Similar analyses on other gene expression concentration–response data would be valuable to confirm our results. Another promising approach, which is not considered in this work, is model averaging. It would be interesting to analyze the influence of the different approaches and parameters on target dose estimation.

ACKNOWLEDGMENTS

This work has received funding from LivSysTransfer (BMBF: Project Number 031L0119) and was partially supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project P2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation: Project Number 427806116). The authors would like to thank the anonymous reviewers for their numerous helpful comments.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The original data that support the findings of this study are openly available in ArrayExpress as stated by Krug et al. (2013) (https://doi.org/10.1007/s00204-012-0967-3).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Julia C. Duda https://orcid.org/0000-0002-3894-0124
Franziska Kappenberg https://orcid.org/0000-0001-8066-5333
Jörg Rahnenführer https://orcid.org/0000-0002-8947-440X

REFERENCES

Abelson, R. P., & Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis general theory and the case of simple order. *The Annals of Mathematical Statistics*, *34*(4), 1347–1369.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bornkamp, B., Bretz, F., Dette, H., & Pinheiro, J. (2011).. Response-adaptive dose-finding under model uncertainty. *The Annals of Applied Statistics*, *5*(2B), https://doi.org/10.1214/10-aoas445

Bornkamp, B., Pinheiro, J., & Bretz, F. (2009). MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, *29*(7), https://doi.org/10.18637/jss.v029.i07

Bretz, F., Pinheiro, J. C., & Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, *61*(3), 738–748.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12–19.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.

Cotariu, D., Zaidman, J. L. (1991). Developmental toxicity of valproic acid. *Life Sciences*, *48*(14), 1341–1350.

European Medicines Agency (2015). *Qualification opinion of mcp-mod as an efficient statistical ethodology for model-based design and analysis of phase II dose finding studies under model uncertainty*. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-qualification-opinion-mcp-mod-efficient-statistical-methodology-model-based-design-analysis_en.pdf.

Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., & Martin, M. T. (2016). tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics*, btw680, https://doi.org/10.1093/bioinformatics/btw680

Food and Drug Administration. (2016). *Determination letter*. https://www.fda.gov/media/99313/download.

Genton, P., Semah, F., & Trinka, E. (2006). Valproic acid in epilepsy. *Drug Safety*, *29*(1), 1–21.

House, J. S., Grimm, F. A., Jima, D. D., Zhou, Y.-H., Rusyn, I., & Wright, F. A. (2017). A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. *Frontiers in Genetics*, *8*, 168.

Kappenberg, F., Grinberg, M., Jiang, X., Kopp-Schneider, A., Hengstler, J. G., & Rahnenführer, J. (2021). Comparison of observation-based and model-based identification of alert concentrations from concentration–expression data. *Bioinformatics*, *37*(14), 1990–1996.

Krug, A. K., Kolde, R., Gaspar, J. A., Rempel, E., Balmer, N. V., Meganathan, K., Vojnits, K., Baquié, M., Waldmann, T., Ensenat-Waser, R., Evans, R. M., Julien, S., Kortenkamp, A., Hescheler, J., Hothorn, L., Bremer, S., van Thriel, C., Krause, K.-H., Hengstler, J. G., Rahnenführer, J., Leist, M., & Sachinidis, A. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: A transcriptomics approach. *Archives of Toxicology*, *87*(1), 123–143.

O'Quigley, J., Iasonos, A., & Bornkamp, B. (2017). *Handbook of methods for designing, monitoring, and analyzing dose-finding trials*. Handbooks of Modern Statistical Methods, CRC Press.

Pinheiro, J., Bornkamp, B., & Bretz, F. (2006). Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics*, *16*(5), 639–656.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Schorning, K., Bornkamp, B., Bretz, F., & Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, *35*(22), 4021–4040.

Thomas, N., Sweeney, K., & Somayaji, V. (2014). Meta-analysis of clinical dose–response in a large drug development portfolio. *Statistics in Biopharmaceutical Research*, *6*(4), 302–317.

Ting, N. (2006). *Dose finding in drug development*. Springer Science & Business Media.

Wheeler, M. W., & Bailer, A. J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*, *16*(1), 37–51.

Xun, X. & Bretz, F. (2017). The MCP-Mod methodology: Practical considerations and the dose finding r package. In J. O'Quigley, A. Iasonos, & B. Bornkamp (Eds.), Handbook of methods for designing, monitoring and analyzing dose-finding trials (pp. 205-227). Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Duda, J.C., Kappenberg, F., & Rahnenführer, J. (2022). Model selection characteristics when using MCP-Mod for dose–response gene expression data. *Biometrical Journal*, *64,* 883–897. https://doi.org/10.1002/bimj.202000250

APPENDIX

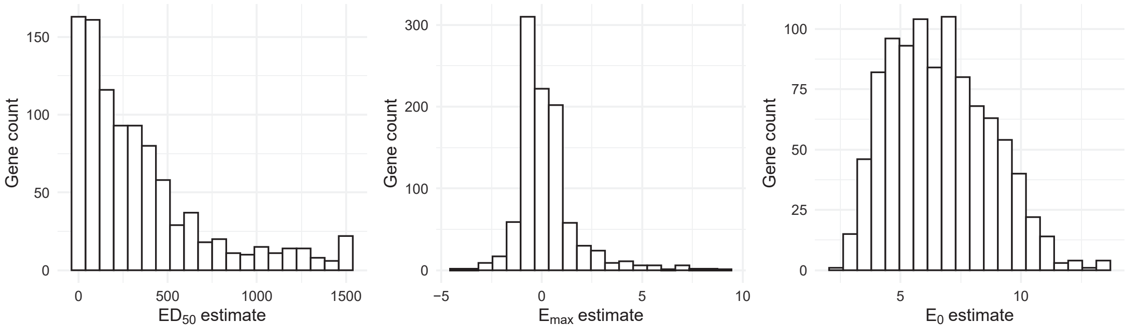


FIGURE A1 Barplots of parameter estimates of the $E_{\max}$ model for genes where it was selected as winner model by the AIC

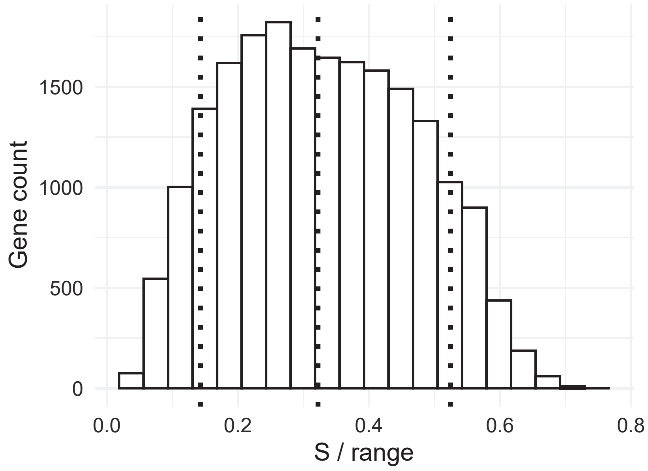


FIGURE A2 Barplot of S to range ratio for genes with an FDR adjusted signifMCP in Analysis I. The vertical dotted lines are at $q(0.1) = 0.1427$, $q(0.5) = 0.3222$, and $q(0.9) = 0.5245$

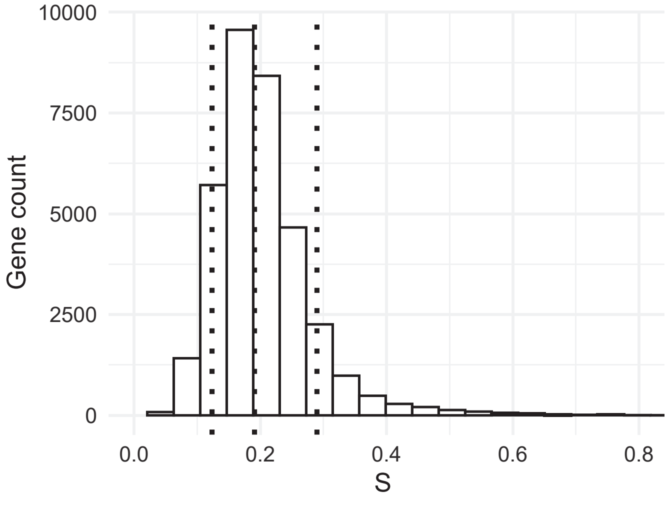


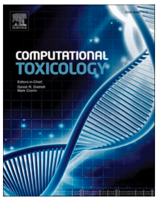
FIGURE A3 Barplot of S for nonsignificant genes in Analysis I. The vertical dotted lines are at $q_{\text{null}}(0.1) = 0.1236$, $q_{\text{null}}(0.5) = 0.1909$, and $q_{\text{null}}(0.9) = 0.2898$

Article 2

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology



td2pLL: An intuitive time-dose-response model for cytotoxicity data with varying exposure durations

Julia Duda [a,*], Jan G. Hengstler [b], Jörg Rahnenführer [a]

[a] *Department of Statistics, TU Dortmund University, Vogelpothsweg 87, Dortmund 44227, Germany*
[b] *Department of Toxicology, TU Dortmund University, Leibniz Research Centre for Working Environment and Human Factors, Ardeystr. 67, Dortmund 44139, Germany*

ARTICLE INFO

Keywords:
Exposure duration-concentration-response
Time-dose-response
Dose-response
Exposure-duration
Cytotoxicity
Statistical modelling

ABSTRACT

Statistical modeling approaches for dose-response or concentration-response analyses are often required in toxicological applications, especially for cytotoxicity assays. By fitting a concentration-response curve, one can derive target concentrations, such as the EC_{50}. In practice, concentration-response data for different exposure durations might be available and the target concentration for each or some exposure duration(s) are of interest. In this work, we propose a statistical modeling approach that improves the precision of the target concentration estimation at a given exposure duration by extrapolating the concentration-response data from other exposure durations. The method further enables target concentration estimation at exposure durations that were not conducted in the experiment. For practitioners, the proposed model yields additional complexity compared to the simple approach of a single concentration-response curve for all exposure durations. It would only be used if it improves the estimation of the target concentration compared to the simple approach. We propose a two-step pipeline to decide between using the complex and the simple approach to result in a precise target concentration estimation.

The methods were evaluated using a simulation study and a real data set. The models are accessible for practitioners through the R package td2pLL.

1. Introduction

In cytotoxicity assays, concentration-response curves help to understand the functional relationship between exposure of cells in the culture medium and viability. To conduct an assay, in addition to the concentration, the exposure duration of the cells should also be set. Gu et al. [1] analyzed the relevance of the exposure or incubation duration on cytotoxicity in primary human hepatocytes (PHH), in which exposure durations of 1 or 2 days are normally administered [2]. They showed a clear influence of exposure duration on EC_{50} values of a set of 30 compounds incubated for 1, 2 and 7 days by separately fitting concentration-response curves for each compound and incubation duration. The selection of one or more exposure duration(s) is therefore relevant for cytotoxicity assays as it affects the target concentration estimation. However, there is no clear guideline for selecting exposure durations for cytotoxicity tests. The guidance document of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), for example, states that longer exposure durations tend to enhance the sensitivity of a test and propose to use exposure durations of at least one cell cycle [3]. This usually leads to recommended exposure durations of 24, 48, or 72 h, depending on the test. In their guidance document, the ICCVAM also referred to Riddell et al. [4], who first addressed the exposure duration question for cytotoxicity tests. They found large differences in the toxicity of compounds between 48 h and 72 h exposure duration. Possible mechanistic explanations are that some substances damage cell membranes while others affect DNA replication or cell division. The toxic effects of the former can be observed after a short time and the latter only after longer exposure durations, as the cell is only affected during certain phases of the cell cycle [5]. Consequently, if the toxicological mechanism of a compound is unknown, one should consider various exposure durations. If toxicity can be measured for short exposure durations, it is of further interest to investigate the exposure duration dependency.

An exposure duration-concentration-response (ECR) model could facilitate answering questions such as: What is the hypothetical limit target concentration for infinitely long exposure durations? When does increasing the exposure duration cease to decrease the target concentration? The mechanistic motivation of the proposed ECR model helps to answer questions relating to dependency between the exposure duration

* Corresponding author.
E-mail addresses: duda@statistik.tu-dortmund.de (J. Duda), hengstler@ifado.de (J.G. Hengstler), rahnenfuehrer@statistik.tu-dortmund.de (J. Rahnenführer).

https://doi.org/10.1016/j.comtox.2022.100234

Available online 1 June 2022

Nomenclature

Abbreviations

td2pLL	time-dose two-parameter log–logistic
EC_{50}	effective concentration of half maximal effect
PHH	primary human hepatocytes
ICCVAM	Interagency Coordinating Committee on the Validation of Alternative Methods
ECR	exposure duration-concentration-response
4pLL	four-parameter log–logistic
2pLL	two-parameter log–logistic
RSS	residual sum of squares
AMAFC	average mean absolute $\log_2$ fold-change
APAP	Acetaminophen
ASP	Aspirin
BOS	Bosentan
BRP	Buspirone
BUSF	Busulfan
CBZ	Carbamazepine
CHL	Chlorpheniramine maleate
CLON	Clonidine
DFN	Diclofenac
DMSO	Dimethyl sulfoxide
ETOH	Ethanol
FAM	Famotidine
GLC	Glucose
HYZ	Hydroxyzine
INAH	Isoniazid
KC	Ketoconazole
LAB	Labetalol
LEV	Levofloxacin
MEL	Melatonin
MePA	Methylparaben
NAC	Acetylcysteine
NFT	Nitrofurantoin
NIM	Nimesulide
PHB	Phenylbutazone
PMZ	Promethazine
PPL	Propanolol
RIF	Rifampicin
TSN	Triclosan
VITC	Vitamin C
VPA	Valporic acid

and the target concentration. The hypothetical target concentration EC_{50} for an infinitely long exposure duration is just one parameter of the model. The parameter and the answer to the question are therefore available as soon as the model is fitted. ECR modeling could solve the issue of exposure-duration selection for concentration-dependent cytotoxicity testing.

Fitting an ECR model has two main advantages over fitting separate concentration-response curves for each exposure duration:

1. The target concentration estimation is not restricted to the exposure durations conducted in the experiment. As the ECR model fits a surface over the combinations of concentration and exposure duration, it predicts a viability response for each possible combination including those that are not conducted in the experiment. ECR model fitting thus yields a target concentration estimation that is less sensitive to the selection of exposure durations than fitting separate concentration-response curves at each exposure duration.
2. A benefit of an ECR model is the increased statistical precision in target concentration estimation. If one fits a one-dimensional concentration-response curve for each exposure duration separately, only the data that belong to the respective exposure duration are used for each fit. When fitting an ECR model, only a single fit is calculated using the data from all exposure durations. This increases the sample size, the precision in estimating the model parameters and, in turn, the target concentration estimation.

These two benefits of the ECR model over separate concentration-response curve fitting at each exposure duration underline the attractiveness of the ECR modeling approach. More recent works highlighted the need (Morin et al. 2018) and efforts (Focke et al. 2017; Serra et al. 2020) to overcome suboptimal separate concentration-response curve fitting when concentration-response relationships at different exposure durations are to be analyzed. These are discussed in Section 2.2. Two-dimensional ECR modeling has key advantages in comparison to separate concentration-response modeling when different exposure durations are conducted in the experiment and current research seeks for good ECR models.

The validity of the model assumptions is crucial for the major improvement of ECR modeling over separate concentration-response fitting. We addressed this considering the mechanistic motivation of the specific ECR model, proposed here in Section 2.1. In addition to the mechanistic motivation, the application of the ECR model on a real ECR cytotoxicity data set is presented in Section 2.4, which supports the model assumptions. The benefits of our model are described, for which the assumptions and possible modifications are justified as far as the limited size of the available data allowed.

To understand Section 2.4, a brief explanation of the concept of overfitting may be required for readers unfamiliar with this term. Generally, a model with many parameters can fit to a data set more closely than a model with fewer parameters. A more flexible model, however, is subject to a high risk of overfitting the data, i.e. it even fits to small deviations in the data that are due to noise. When *overfitting* is strong, the resulting model would perform poorly on new data if the experiment is repeated. The overfitting effect is always present, but less strong or negligible for large data sets and less parameters. For small data sets and many parameters its effect is stronger. Hence, the number of parameters must be accounted for, i.e. the flexibility of a model, when comparing how closely a model fits the data as a measure of model performance. Otherwise, a model with more parameters will automatically fit the data more closely than a model with fewer parameters, but the seemingly better performance in the sense of a closer fit can be a result of overfitting.

In addition to the proposed ECR model, we propose a two-step pipeline to automatically distinguish between experimental data for which fitting an ECR model is beneficial and where it might be unnecessary complicated. It is possible that concentration-response data are available for different exposure durations, but all exposure durations are rather long so that the target concentration (such as the EC_{50}) does not differ between these exposure concentrations. In this case, it would be unnecessarily complicated and would not yield better target concentration estimation when fitting the proposed ECR model. For these cases, it is preferred both in terms of a better target-concentration estimations and simplicity of the model to treat the data (after normalization with exposure duration-wise control) as if there is only one exposure duration. A simple concentration-response curve should be fitted using the data of all exposure durations. To account automatically for such a scenario, we additionally propose a statistical two-step procedure. In step 1, it is decided objectively whether the EC_{50} differs between different exposure durations. If so, the ECR model is fitted in step 2. If not, a single concentration-response curve is fitted in step 2. The proposed two-step pipeline helps practitioners to fit an ECR model only on experimental data where it is beneficial and necessary.

To quantify and compare the potential benefit of fitting the proposed ECR model and the two-step pipeline, we conducted a simulation study based on real cytotoxicity data of Gu et al. [1], which is presented in Section 2.5. We compared the new model and pipeline to fitting separate concentration-response curves, as well as fitting a single concentration-response curve to all exposure durations. The latter approach effectively ignores that there are different exposure durations. The simulation study accounts for many scenarios, such as different exposure duration effects on the target concentration EC_{50} (no, small and large effect) and different levels of background noise (small, medium, large) in the data.

A common bottleneck for implementing sound theory into practical usage is that sophisticated statistical methods are often difficult to grasp for practitioners with less profound mathematical background or there is a lack of software availability. These drawbacks can prevent theoretically well-established methods from being used in practice. Hence, simplicity and good interpretation of the proposed ECR model is an advantage compared to other works in this field, as discussed in Section 2.1. To overcome problems with practical implementation and calculations of our proposed ECR model and methods, they were made available through an open source R software [6] package td2pLL via GitHub (https://github.com/jcduda/td2pLL/). We therefore eased the usage of our methods for practitioners both by making our model easy to interpret and by providing a software implementation.

In this study, we addressed the question as to whether there is a benefit in using the proposed ECR model and the two-step pipeline when estimating target concentrations such as the EC_{50} for cytotoxicity data, where different exposure durations are conducted in the experiment. The methods have intrinsic advantages compared to fitting separate concentration-response curves for each exposure duration and they improve target concentration estimation, which was demonstrated in a simulation study based on real data.

We note that in the mathematical literature, *time-dose-response* model is the general term and hence motivates the chosen name td2pLL (time-dose two-parameter log–logistic) model. However, we used the biologically correct term *exposure duration-concentration-response* (ECR) model throughout this manuscript.

2. Materials and methods

2.1. Exposure duration-concentration-response model

In this section, we explain how the proposed ECR model is an extension of a commonly used concentration-response curve. A detailed explanation if given of how the parameters of the model can be interpreted in terms of toxicological research interests.

For cytotoxicity assays, the most popular model used for describing concentration-response curves is the sigmoidal four-parameter log–logistic (4pLL) model ([7–9]):

$$f\left(x\right) = E_0 + E_{\max}\frac{x^h}{x^h + EC_{50}^h}, \tag{1}$$

where x is the concentration and f(x) the viability in %. The Hill or slope parameter, h, represents the steepness of the curve, with a greater value indicating a steeper curve. EC_{50} is the concentration where the half-maximal effect is reached, i.e. half of $E_{\max}$. For clarity, note that the other well-known parametrization ([7]) of the 4pLL model is

$$f\left(x\right) = c + \frac{d-c}{1+\exp\{b(\log(x)-\log(e))\}} = c + \frac{d-c}{1+\left(\frac{x}{e}\right)^b}$$

where d is the upper asymptote, c is the lower asymptote, e is the EC_{50} and b is h. For viability assays, one can assume that after normalization with respect to the raw mean response at the control, the viability is 100 [%] at the control, and for large enough concentrations, the viability tends towards 0[%]. Mathematically, this is equivalent to setting $E_0 = 100$ and $E_{\max} = -100$ in Eq. (1), yielding:

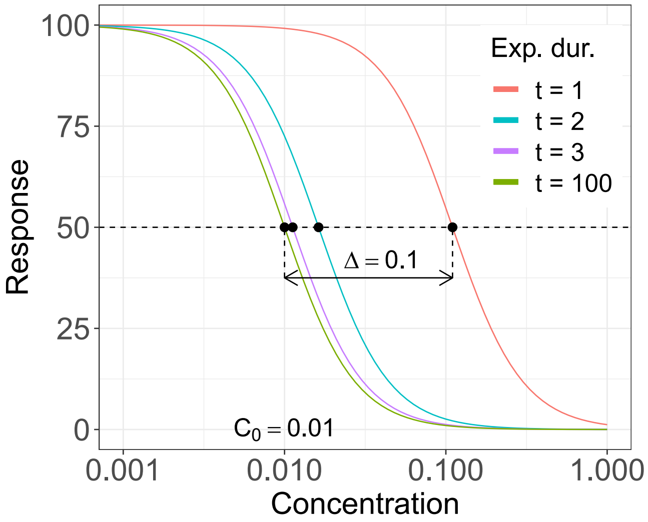


Fig. 1. Concentration-response curves of a td2pLL model at different exposure durations t with parameter $\Delta = 0.1$. Black symbols on the dotted line show how the $EC_{50}(t)$ moves dependent on t.

$$f\left(x\right) = 100 - 100\frac{x^h}{x^h + EC_{50}^h}. \tag{2}$$

Hence, for cytotoxicity one can reduce the 4pLL model to the 2pLL model, as only two parameters, EC_{50} and h, remain to be estimated. In cytotoxicity assays it can occur that the viability does not reach 0%. For an initial development of an ECR model, we restrict ourselves to the reasonable assumption of 0% viability at large concentration. Examples of 2pLL curves are presented in Fig. 1.

The 2pLL concentration-response model depends only on the concentration. To incorporate exposure duration, the EC_{50} parameter is modelled to be dependent on the exposure duration, t, by setting $EC_{50} = EC_{50}(t)$. This is intuitive because, e.g., for a longer exposure duration, one expects a smaller concentration to yield 50% viability. How exactly the $EC_{50}(t)$ changes according to the exposure duration, t, needs to be formulated mathematically. We therefore applied a modified version of Haber's law [10] following Miller et al. [11]. Originally, Haber postulated that a lethal effect of a compound is determined by multiplying the concentration of the compound and the exposure duration, i.e. effect $=$ concentration $\cdot$ exposure duration. For exposure duration $t = 1$, we denote the lethal effect for which half of the cells die, by $\widetilde{\Delta}$. By definition, the corresponding concentration is the EC_{50}:

$$\widetilde{\Delta} = EC_{50}\cdot t.$$

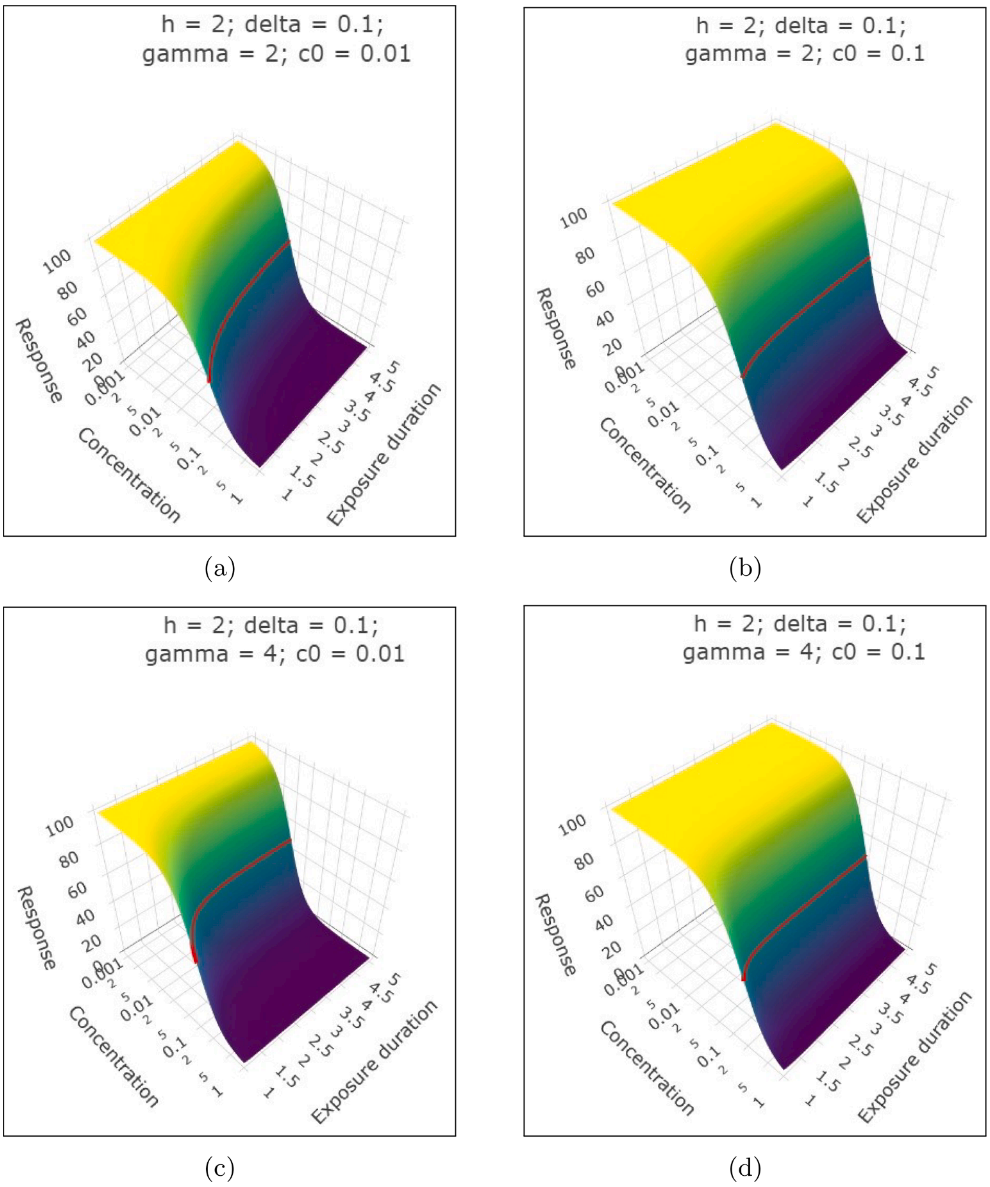
Miller et al. [11] further introduced a parameter C_0 as a lower limit for $EC_{50}(t)$. The parameter C_0 can be interpreted as the EC_{50} at an infinitely large exposure duration. This leads to the equation

$$\widetilde{\Delta} = (EC_{50} - C_0)\cdot t.$$

Lastly, concentration and exposure duration might not change the lethal effect equally when increased or decreased by one unit. To capture this, different exponents, α and β, were added in the model:

$$\widetilde{\Delta} = (EC_{50} - C_0)^{\alpha}\cdot t^{\beta}.$$

Solving the equation for EC_{50} yields:



(a)

(b)

(c)

(d)

Fig. 2. Explanation on how to interpret the modeled influence of exposure duration on EC_{50} (red line) when fitting the proposed exposure duration-concentration-response (ECR) model with four ((a)-(d)) example parametrizations of the proposed ECR model, td2pLL. They have the same steepness parameter $h = 2$ and the same Δ parameter. Δ is the difference in EC_{50} values at exposure duration $t = 1$ and hypothetical exposure duration $t \to \infty$, where $EC_{50}(t \to \infty) = C_0$. Comparing the top row ((a), (b)) with the bottom row ((c), (d)), γ increases, which explains the stronger change of the EC_{50} at short exposure durations. In the right column ((b), (d)), the EC_{50} does not appear to change much between different exposure durations compared to the left column ((a), (c)). This is because the concentration is visualized on a logarithmic scale and the fold-change between $C_0 + \Delta = 0.2$ ($= EC_{50}$ at exposure duration $t = 1$) and $C_0 = 0.1$ is only $0.2/0.1 = 2$, compared to 11 in (a) and (c).

$$\underbrace{\tilde{\Delta}^{1/\alpha}}_{:=\Delta} t^{-\beta/\alpha} + C_0 = EC_{50}.$$

Mathematically, the above parametrization does not lead to a unique solution when fitting the model to an appropriate data set. A reparametrization circumvents this technical problem, yielding the final dependency of the EC_{50} on the exposure duration:

$$\overset{\gamma=\beta/\alpha}{\Leftrightarrow} \Delta \cdot t^{-\gamma} + C_0 = EC_{50}(t).$$

Note that as soon as the parameters, Δ, γ and C_0 are known/estimated, the $EC_{50}(t)$ can be derived/estimated directly at any exposure duration t of interest using the above formula. There is no restriction to exposure durations conducted in the experiment. By plugging $EC_{50}(t)$ into the 2pLL model, the new td2pLL ECR model is obtained:

$$f\left(t, x\right) = 100 - 100\frac{x^h}{x^h + (\Delta \cdot t^{-\gamma} + C_0)^h}. \tag{3}$$

Examples of the model are visualized in Fig. 2. The interpretation of the slope parameter, h, remains as in the standard concentration-response 2pLL model: A larger h leads to a steeper curve alongside the concentration. The parameters that define how the EC_{50} changes for different exposure durations are C_0, γ, and Δ and can be interpreted as follows:

- C_0 is the limit EC_{50} for a hypothetical, infinitely long exposure duration. For increasing exposure duration t, $EC_{50}(t)$ approaches C_0 (if $\gamma > 0$, cf. Figs. 2 and 1).
- $\gamma > 0$ indicates when the $EC_{50}(t)$ changes most between different exposure durations. A larger γ indicates a stronger initial change of the EC_{50} followed by quickly reaching a plateau in the EC_{50}. This means that for shorter exposure durations the EC_{50} differs markedly and for larger exposure durations the EC_{50} does not change much as it is already close to its limit C_0. A smaller γ indicates a more balanced change of the EC_{50} between exposure durations. This means that the EC_{50} differs considerably even for two comparatively long exposure durations. In other words, the EC_{50} reaches a plateau less quickly, i.e. at even longer exposure durations compared to larger $\gamma > 0$.
- $\gamma < 0$ is the unintuitive case that the EC_{50} increases when the exposure duration increases. This increase could be valid if the substance

is beneficial instead of toxic. Mathematically, the model contains no constraints on γ, but practically this case is typically irrelevant.

- Δ is the difference in EC_{50} values at exposure duration $t = 1$ and exposure duration $t = \infty$, i.e. the limit EC_{50} value for infinitely large exposure durations, which is C_0. Δ hence indicates the range of the EC_{50} values for exposure durations $t \geqslant 1$ (cf. Fig. 1). The EC_{50} at exposure duration $t = 1$ is always $\Delta + C_0$. An example of the interpretation of Δ and C_0 is described below. Mathematically, $EC_{50}(t = 1) - EC_{50}(t \to \infty) = (\Delta + C_0) - C_0 = \Delta$.

In the following we explain the interpretation of the model parameters C_0 and Δ using examples (Figs. 2 and 1). When considering fold changes in the EC_{50} estimates between the exposure duration $t = 1$ and the hypothetical exposure duration, $t \to \infty$, both estimates of Δ and C_0 should be conducted simultaneously. For example, given the estimates $C_0 = 0.01$ and $\Delta = 0.1$, the fold change is 11. This can be derived directly from the model parameters by

$$EC_{50}(t = 1)/EC_{50}(t \to \infty) = (C_0 + \Delta)/C_0 = (0.01 + 0.1)/0.01 = 11.$$

However, if the estimated EC_{50} at an infinitely long exposure duration is $C_0 = 0.1$ while still $\Delta = 0.1$, then there is only a fold change of $(0.1 + 0.1)/0.1 = 2$ in the EC_{50} value between exposure durations $t = 1$ and $t \to \infty$. In summary, the proposed ECR model, td2pLL, has a clear derivation as it naturally extends the commonly used sigmoidal concentration-response using Haber's law [10].

2.2. Comparison to other approaches

Efforts on modeling ECR data have been increasing lately. Focke et al. [12] used a similar approach to ours, but the exposure duration-response relationship was modeled with a sigmoidal function and the concentration component was added according to Haber's law. This resulted in a different ECR model. There was no software implementation provided by Focke et al. [12]. In the work of Schüttler et al. [13], an ECR model for toxicogenomic data was proposed where the concentration-response relationship was a 4pLL model and the exposure duration was also incorporated using the EC_{50} parameter. However, the dependence of the EC_{50} parameter was modeled using a non-monotone logarithmic Gaussian function. The authors justified this type of dependency with its good empirical fit to toxicogenomic data. For cytotoxicity data, however, a non-monotone exposure duration dependency is less realistic. A software implementation was provided by Schüttler et al. [13] through an R-package. In the R-package TinderMIX introduced by Serra et al. [14], ECR toxicogenomic data were modeled using linear regression with polynomials with a maximal order of three. One obvious limitation is that these models might predict unrealistic (e.g. negative) response values for exposure duration-concentration settings that are not close to the exposure duration-concentration settings of the experiment. Furthermore, the model parameters are not easily interpretable if quadratic terms or terms of a higher order are included. With td2pLL, we fill a scientific gap by providing a mechanistically motivated, easily interpretable model for cytotoxicity data that is embedded in an R-package to ease its application for practitioners.

2.3. Two-step pipeline for exposure duration-concentration-response data

Even when exposure duration-resolved and concentration-resolved toxicity data are available, they might not exhibit a clear exposure duration-effect. One possible reason is that the experimentally chosen exposure durations all lie within a range where the exposure duration-effect is already saturated and the EC_{50} no longer changes. For such cases, it is unnecessarily complicated and not appropriate to use the proposed ECR model. Instead, ignoring the information on exposure duration and fitting a single concentration-response curve would be considered reasonable. To automatically decide which case applies to the experimental data, we propose a two-step pipeline. Step 1 involves deciding statistically if the exposure duration has an influence on the EC_{50}. If it does, an ECR model is used for fitting in step 2. If not, a single, one-dimensional 2pLL concentration-response curve is fitted on all data in step 2 and the information on exposure duration is ignored (Fig. 3).

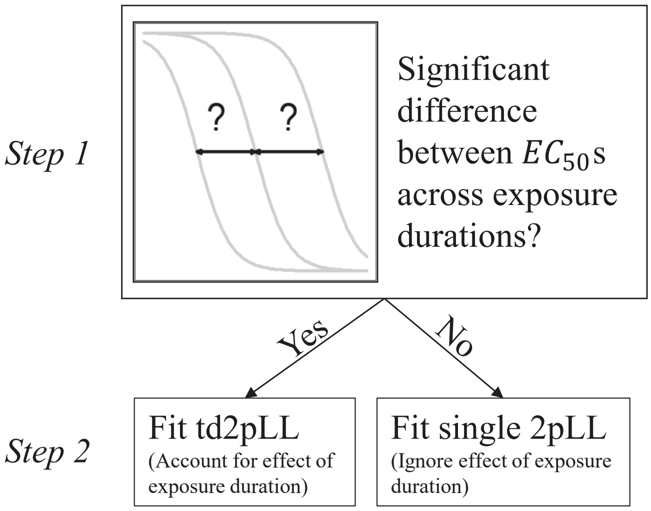

Step 1
?
?
Significant difference between EC_{50}s across exposure durations?
Yes
No
Step 2
Fit td2pLL
(Account for effect of exposure duration)
Fit single 2pLL
(Ignore effect of exposure duration)

Fig. 3. The proposed two-step pipeline decides in an objective, statistical manner if the two-dimensional model should be used to model both concentration- and exposure duration-dependency, or if it suffices to use a one-dimensional model that only includes the concentration.

For step 1, an ANOVA-based test can be used. Statistically, the hypothesis that the EC_{50} is the same across all considered exposure durations t_i is tested against the hypothesis that it is different for any pair of values t_1 and t_2. Here, a typical signal plus noise model

$$y_{ijk} = f(t_i, x_j) + \varepsilon_{ijk} \tag{4}$$

with $i = 1, \dots, k$ exposure duration levels, $j = 1, \dots, l$ concentration levels, n_{ij} replicates at exposure duration-concentration setting (t_i, x_j), and corresponding noise $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ is assumed.

The two nested models Q_0 and Q_1 are then compared. Model Q_0 is a regular 2pLL concentration-response model that is fitted to the data, where exposure duration is completely neglected. Model Q_1 allows for each exposure duration an individual EC_{50} parameter. The remaining hill parameter h remains shared across exposure duration levels. Given the assumption of normally distributed errors, the (nested) ANOVA statistic is

$$F = \frac{(RSS_{Q_0} - RSS_{Q_1})/(\eta - \nu)}{RSS_{Q_1}/\nu}, \tag{5}$$

where η are the residual degrees of freedom for Q_0, ν are the residual degrees of freedom for Q_1, and RSS is the residual sum of squares. The test rejects the hypothesis that Q_0 is the true model at significance level α, if the observed value for F is greater than $F(1 - \alpha, \eta - \nu, \nu)$, the $1 - \alpha$ quantile of an F-distribution with $\eta - \nu$ and ν degrees of freedom. A rejection of the null hypothesis that Q_0 is the true model means that the hypothesis that there is no exposure duration-dependency for the EC_{50} is rejected.

Hence, rejecting the null hypothesis in step 1 leads to fitting the ECR model in step 2. If the test does not reject the null hypothesis in step 1, information on exposure duration is ignored and a single 2pLL concentration-response model is fitted in step 2. The significance level α must be chosen in advance.

The two-step pipeline provides an objective framework to make a decision regarding the data as to whether it is beneficial to account for possible differences in target concentrations across different exposure

Table 1
Overview of the simulation study setup that compares the precision of the EC_{50} estimation between the proposed method (Two-Step) and other methods, based on various simulated scenarios that use real cytotoxicity data from Gu et al. [1].

Aim	Comparing the performance of the proposed exposure duration-concentration-response modeling approach (Two-step pipeline) for toxicity assays with respect to target concentration estimation precision with other approaches.
Data Generating Mechanism	Based on data from Gu et al. [1] with models $M \in \{M_0, M_1, M_2\} =: \mathscr{M}$ where - M_0 is the scenario where the EC_{50} does not change across exposure durations, - M_1 is the scenario where the EC_{50} changes moderately across exposure durations, - M_2 is the scenario where the EC_{50} changes a lot between exposure durations. -Experimental design -$k \in \{3, 4\} =: \mathscr{K}$ different exposure durations -$n_{\text{obs}} \in \{72, 216\} =: \mathscr{N}_{\text{obs}}$ observations -(Normal) noise $N \in \{N_1, N_2, N_3\} =: \mathscr{N}_{\text{oise}}$
Estimands	EC_{50} (Concentration causing a response that is 50 percent of the maximum achievable)
Methods	-**Two-Step**: Proposed two-step-procedure with ANOVA pre-test in step 1 and td2pLL ECR model fit or 2pLL concentration-response model fit with ignored exposure duration, respectively, in step 2. -**Always td2pLL** ECR model fit -**Always separate 2pLL** concentration-response model fit per exposure duration -**Always single 2pLL** concentration-response model fit neglecting exposure duration and thus pooling with respect to exposure durations
Performance	-AMAFC: Mean of the absolute $\log_2$ fold change of estimated and true EC_{50} across exposure durations and averaged over simulation replications. A small AMAFC is desirable.

durations by fitting the proposed ECR model, or if it is sufficient to fit a concentration-response curve.

2.4. Data

The data on which our simulation study is based come from Gu et al. [1] where cytotoxicity testing was performed on primary human hepatocytes. These data are publicly available and the publication contains details of the laboratory protocol. The data set contains concentration-response data of 30 compounds. For each compound, 3 biological replicates (donors) are available. For each donor, there are measurements of 3 exposure durations (1, 2 and 7 days) and (including solvent) 6, 7, or 8 concentrations (mostly 6). The concentrations are equidistant on a log-scale with base $\sqrt{10}$. There are 4 or 8 measurements per compound, donor, exposure duration and concentration. Note that the data was normalized: For each compound, the viability values per donor were divided by the respective donor-solvent response value and multiplied by 100. This normalization implicitly assumes that compound, donor and exposure duration only affect the response at concentration 0. Other correlations cannot be accounted for with this typical normalization procedure. More over, the ‘retting’ approach by Kappenberg et al. [15] was applied as an additional pre-processing step, to better justify the assumption on the asymptotes. Lastly, the concentration-range was rescaled to [0,1] to guarantee comparison of the results across compounds.

2.5. Simulation study - setup

2.5.1. Overview

To compare the performance of our method to other statistical analysis approaches, we used a simulation study based on the cytotoxicity data of Gu et al. [1]. The computations are performed with the td2pLL R package (version 1.0.0.). In such a simulation study, one assumes different scenarios of ECR relationships as *true*. Given such a true scenario, data can be generated and the methods are applied on the generated data to estimate a target concentration for the assumed, true ECR scenario. For reliable simulation results, for each scenario, the data generation and method application process was repeated many ($n_{\text{sim}} = 1000$) times. Performance of the methods was assessed by averaging precision of target concentration estimation, $\widehat{EC}_{50}$, across the different simulation scenarios.

An overview of the main components of the simulation study using the ADEMP summarization principle [16] (Aim, Data Generating Mechanism, Estimands, Methods, Performance) is provided in Table 1. In our simulation study, the general *estimand* was the target concentration EC_{50}. To enhance the generalization of the simulation study results on real data, many factors of data generation were crossed with each other to generate a large, robust pool of simulation scenarios.

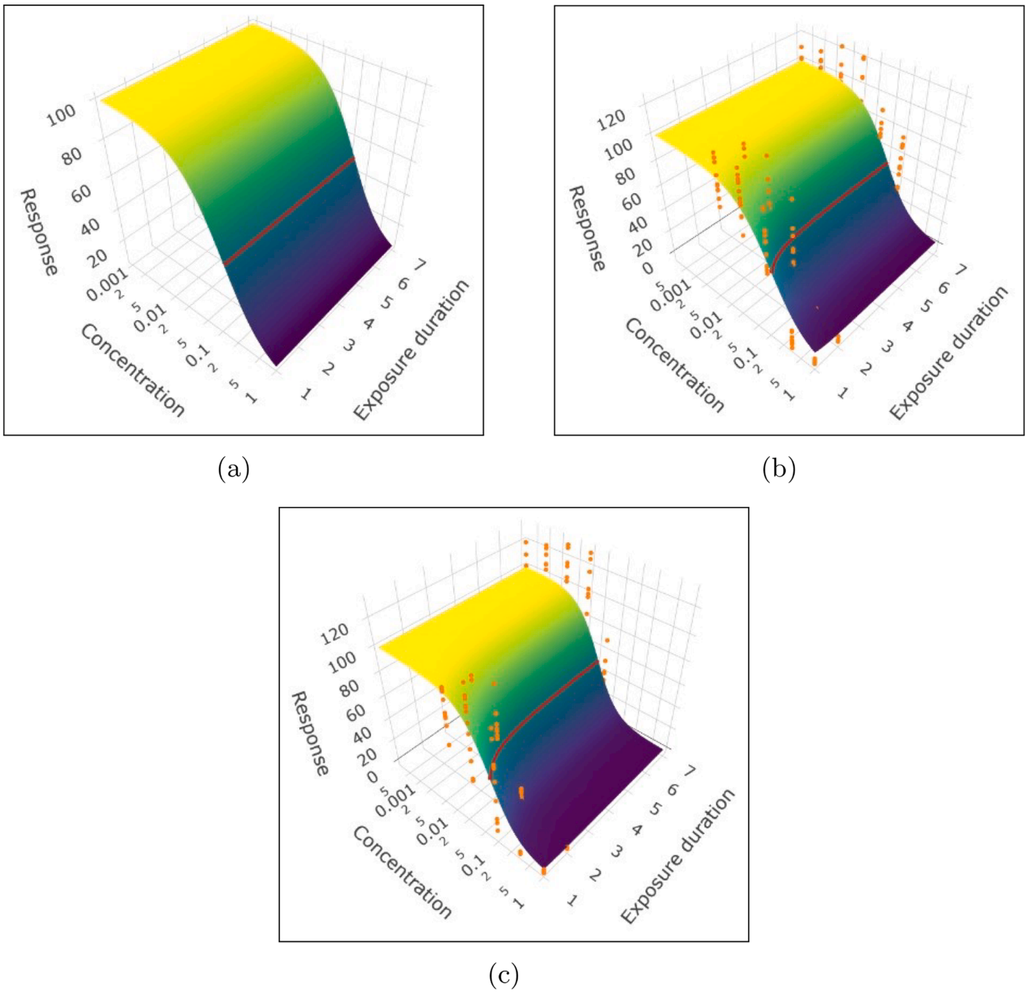
To understand the presentation of the results in Section 3.2, important factors of the data generation should be noted. To mimic the possibility of different magnitudes of influence of the exposure duration on the target concentration in real data, in the simulation study we considered three levels of influence of the exposure duration on the target concentration EC_{50}, namely no (M_0), little (M_1) and strong (M_2) influence. We further accounted for varying levels of background noise in real data by including little (N_1), medium (N_2) and strong (N_3) background noise scenarios in the simulated data. In the following, further detailed explanations on the data generating mechanism, methods, and performance measures are presented.

2.5.2. Details

To generate ECR data for the simulation study, the assumed true model M, the experimental design, and the variability N of the added background noise must be specified. Three models representing different types of exposure duration dependency were used. These models were derived from models fitted on real data. A strong exposure duration dependency means that the EC_{50} differs a lot between different exposure durations. For a strong (M_2) and a weak (M_1) exposure duration dependency, the td2pLL model was fitted to data from two compounds tested by Gu et al. [1], namely chlorpheniramine (CHL) and ethanol, respectively. A model without exposure duration dependency (M_0) was derived from the fitted model M_1 by setting $\gamma = 0$, which removed the influence of the exposure duration on the viability response such that the EC_{50} was the same across all exposure durations (Fig. 4). To enhance comparability, the concentration ranges of the compounds were rescaled to be between 0 and 1.

To cover more aspects of a real experiment in the simulation scenarios, we varied the number of exposure durations in the generated data. Given a true ECR model M from the set $\{M_0, M_1, M_2\}$, the true response values are calculated at either $k = 3$ or $k = 4$ different exposure durations as depicted in Fig. 5. For 3 exposure durations, the selected exposure duration-concentration points (t_i, x_j) agree with the ones selected in the real experiment of Gu et al. [1] for compound CHL. Note that in Gu et al. an effect of the exposure duration was anticipated, such that for the larger exposure durations of 7 (days), lower concentrations were selected. In the simulation, either $n_{\text{obs}} = 72$ or $n_{\text{obs}} = 216$ observations are generated with equal sample size for all exposure duration-concentration combinations (t_i, x_j).

To understand the influence of noise, we added different levels of background noise to the assumed true mean responses. It was assumed that all data points at a specific exposure duration-concentration point



(a)

(b)

(c)

Fig. 4. To increase the representativeness of the simulation study for real data, three different scenarios of effect (none, weak, large) of exposure duration on EC_{50} were included in the simulation. The weak (b, model M_1) and large (c, model M_2) effect are each a td2pLL model fit on real ECR data from Gu et al. [1] of the compounds ethanol and chlorpheniramine. The no-effect model M_0 (a) was derived from the low-effect model M_1 (b), with complete elimination of the exposure duration effect by setting $\gamma = 0$. The models M_0, M_1, and M_2 served as assumed true exposure duration-concentration-response relationships for the simulation study. Exact parameters of the models are: M_0 ($h = 1.26, \Delta = 0.07, \gamma = 0, C_0 = 0.05$), M_1 ($h = 1.26, \Delta = 0.07, \gamma = 2.19, C_0 = 0.05$), M_2 ($h = 1.54, \Delta = 0.06, \gamma = 2.05, C_0 = 0.02$). Note that the controls could not be shown due to the logarithmic scale of the concentration axis.

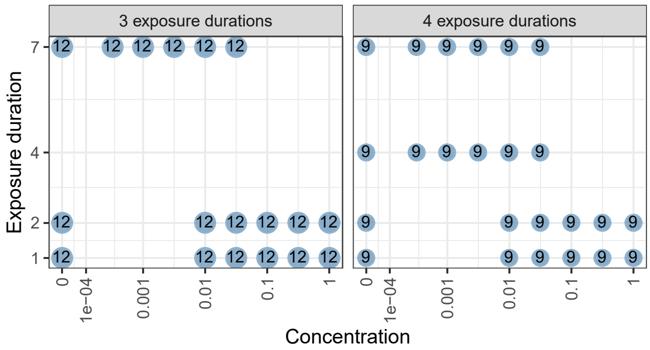


Fig. 5. To increase the representativeness of the simulation study for real data, the number of exposure durations was varied in the simulation. Experimental designs are shown with $k = 3$ (left) and $k = 4$ (right) different exposure durations for the data generating mechanism in the simulation study. Only the cases with $n_{\text{obs}} = 216$ are depicted. For $n_{\text{obs}} = 72$, the replicates at each exposure duration-concentration point were reduced to 4 and 3, respectively. Note that a pseudo-log transformation was used to display the control at concentration-level 0.

(t_i, x_j) have the same, true mean response $f_M(t_i, x_j)$ of model $M \in \{M_1, M_1, M_2\}$. Noise was added as realizations of independent, normally distributed, mean zero noise (cf. Eq. (4)). The noise standard deviation, σ, was chosen based on a linear model of the empirical standard deviations observed in the real data of all compounds in Gu et al. [1]. This data-driven approach yields more realistic simulated data. For each compound, donor and exposure duration-concentration point (t_i, x_j), the standard deviation $\sigma_{i,j}$ of the pre-processed (cf. Section 2.4) cytotoxicity measurements was calculated. Note that this procedure calculated a single standard deviation value from 3 to 4 measurements. A model was fitted to the standard deviations depending on the exposure duration, t_i, and pseudo-log transformed concentrations, x_i (see Figs. 5 and 6). The outcome (standard deviation) was modeled as a linear combination of intercept, concentration, concentration^2, exposure duration and the interaction between concentration^2 and exposure duration. This means that the concentration can have a quadratic effect and exposure duration a linear effect on the standard deviation. Only significant variables (p-value < 0.05) were retained in the model. The resulting model was used to calculate the standard deviation for the normally distributed noise $\varepsilon \sim \mathcal{N}(0, (\sigma_{i,j})^2)$. From ε, realizations were drawn that were added to the response values $f_M(t_i, x_j)$.

We considered 3 scenarios with increasing noise levels. In N_1 (low noise), the generated standard deviations were divided by 2, in N_2 (medium noise) they were left unchanged, and in N_3 (high noise), they were multiplied by 2. This approach of adding background in the simulation study closely mimicked real data background noise and therefore further increased the representativeness of the simulation study for real ECR data.

To compare the approaches considered in this work for fitting ECR data and predicting target concentrations, each approach was applied on each ECR data generating scenario. All $3 \cdot 2 \cdot 2 \cdot 3 = 36$ combinations of data generation in the space $\mathcal{M} \times \mathcal{K} \times \mathcal{N}_{\text{obs}} \times \mathcal{N}_{\text{oise}}$ were crossed with

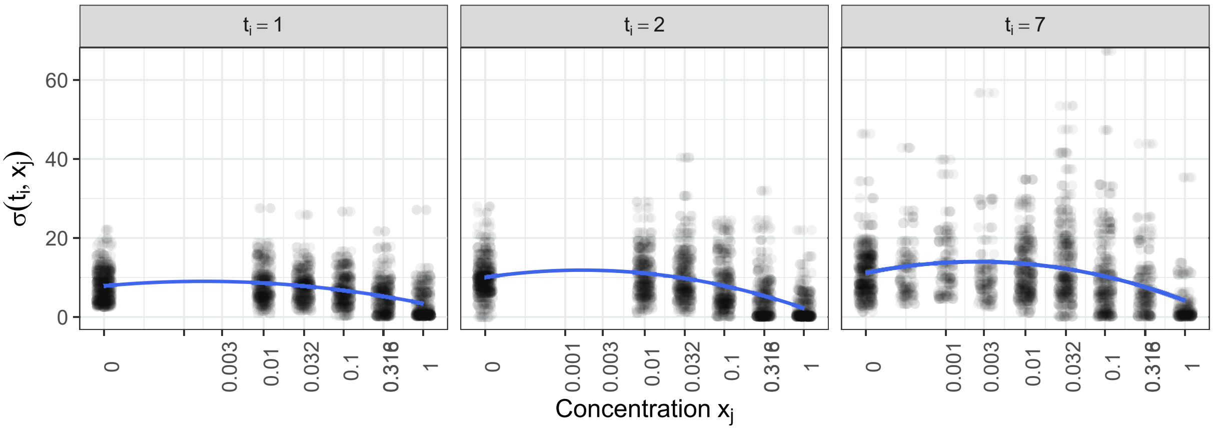


Fig. 6. The model used for determining exposure duration-concentration dependent average noise levels in the simulation study captured properties of real data background noise. Points are empirical standard deviations per compound, donor, concentration and exposure duration in the pre-processed data of Gu et al. [1]. The blue line shows a model fit to these standard errors depending on the exposure duration t_i and the pseudo-log transformed concentration x_j. The model fit was used to select exposure duration-concentration point dependent standard deviations of the noise that was added to the generated data in the simulation. Note that the model captures intuitive noise characteristics, such as a decrease in noise at the maximal concentration.

each method for data fitting (Two-step, always td2pLL, always separate 2pLL, always single 2pLL, cf. Table 1), which yielded $36 \cdot 4 = 144$ simulation scenarios with $n_{\text{sim}} = 1000$ repetitions for each scenario in total. Note that fitting separate 2pLL curves for each exposure duration represented the most basic concentration-response modeling approach for accounting for an exposure duration effect on the target concentration.

For each simulation scenario, the performance of a model fit and associated target concentration estimation was measured using the averaged mean absolute $\log_2$ fold change (AMAFC). For a fixed model and a fixed exposure duration, t_i, $\widehat{EC_{50}}(t_i)$ was the estimated EC_{50} value and $EC_{50}(t_i)$ was the true value. The AMAFC was defined as the mean of the values for $|\log_2(\widehat{ED_{50}}(t_i)/ED_{50}(t_i))|$, where the mean was taken over the exposure durations, t_i, and these means were averaged over all n_{sim} simulation repetitions. A small AMAFC is desirable.

In summary, the resulting simulation was based on real exposure duration-concentration response data of Gu et al. [1] and compared the proposed method regarding estimation precision in target-concentration estimation with several other approaches across many realistic scenarios.

3. Results

3.1. Real data application

To validate the assumptions of our newly proposed model, model fits on real ECR data were analyzed. Validation of the assumptions of our new model is crucial as they are the basis for the benefits of the model, in terms of a potentially improved target concentration estimation for a given exposure duration, by also exploiting data from other exposure durations. We therefore fitted our newly proposed model, the td2pLL model, to pre-processed ECR cytotoxicity data of Gu et al. [1]. We compared the fit of the td2pLL model (approach 1) with separate 2pLL fits for each exposure duration (approach 2). This helped to investigate whether the td2pLL model has valid assumptions for application to cytotoxicity data, thus improving target concentration estimations.

To properly compare the two approaches with respect to the validity of the td2pLL model assumptions, the statistical concept of overfitting is crucial, as explained in the introduction. Approach 2 involved more parameters because for each of the three exposure durations, a new 2pLL model with 2 parameters each was fitted. Also, for each fit only the data of the respective exposure duration were used. For approach one, only 4 parameters were fitted. Hence, approach 2 was more prone to overfitting than approach 1, where the td2pLL was fitted only once to ECR data of all exposure durations and therefore used fewer parameters in total to fit all data of one compound.

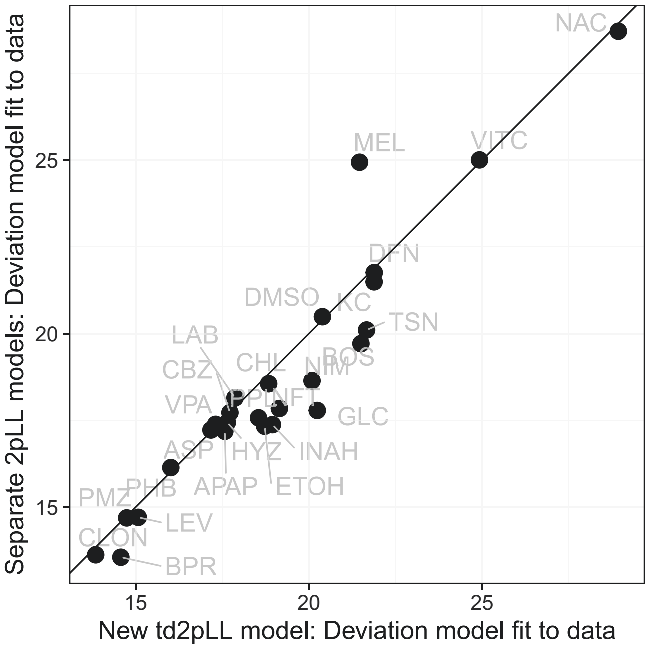


Fig. 7. The newly proposed td2pLL model for ECR data has valid assumptions based on its application to a small cytotoxicity data set, which allows for a more precise target concentration estimation. The comparison of the td2pLL model fits with the classical separate 2pLL model fits per exposure duration for all compounds of Gu et al. [1] based on the deviation between the model fit and the data, which is measured by the standard error (SE) of the residuals. A small SE is desirable, which may however be artificially decreased due to overfitting. The points scatter around the diagonal line, which means that the td2pLL model fits are comparable to those of the separate 2pLL fits. This indicates that the assumptions of the td2pLL model are valid. For compounds BUSF, MePA, RIF and FAM, computational problems in the separate 2pLL fitting occurred and the data points are not shown (cf. Fig. A1). For those compounds, there was no concentration-dependent decline in viability for some exposure durations.

We reduced the overfitting effect in the comparison of the approaches by calculating the standard error (SE) for both approaches for each compound. Note that this measure takes into account the numbers

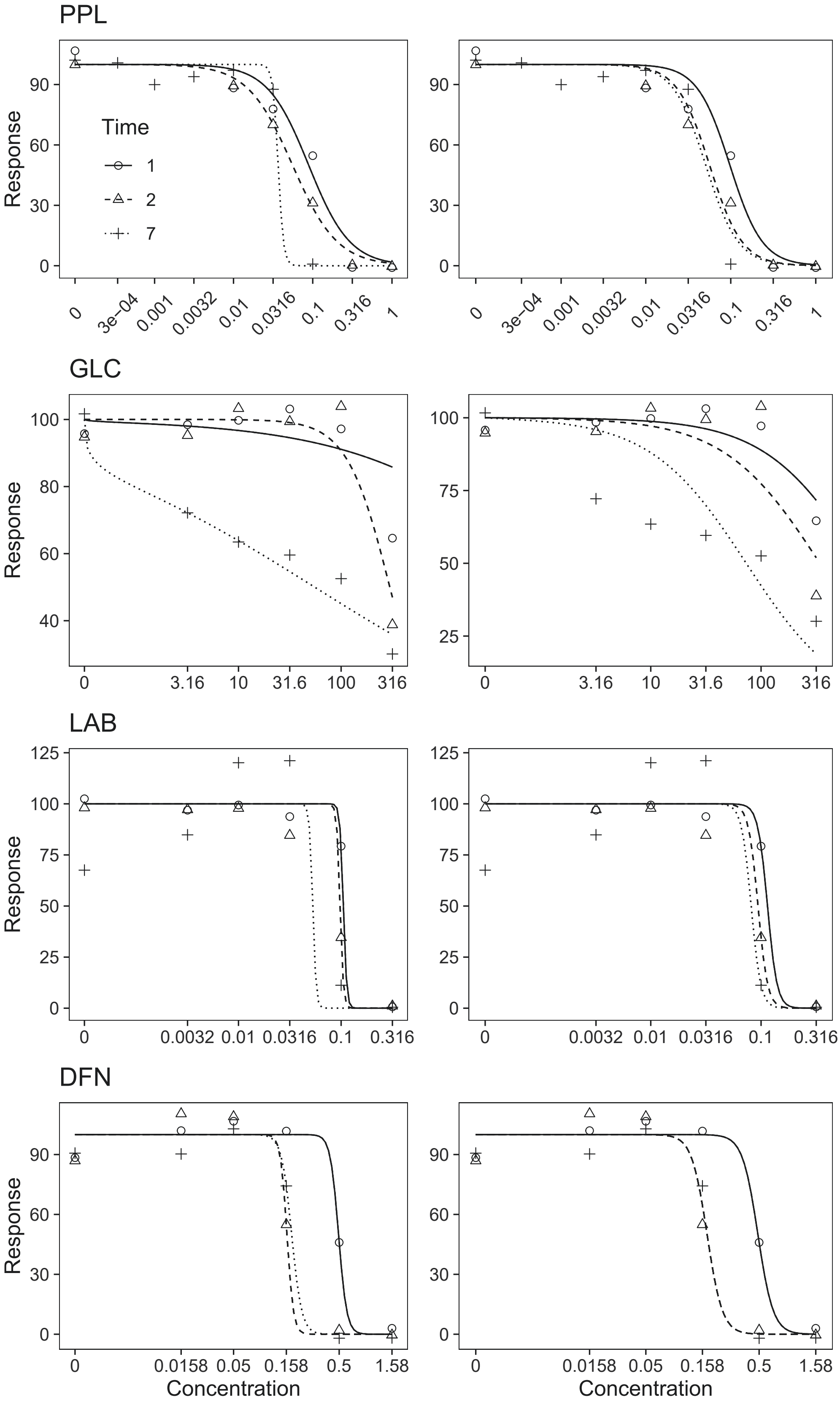


Fig. 8. The newly proposed td2pLL model leads to clear improvements in model fitting for some compounds (LAB, DFN) of [1] and less optimal fits for other compounds (PPL, GLC). The resulting concentration-response curves at exposure durations 1, 2 and 7 (days - legend in upper left plot) are shown for separate 2pLL models (left column) and for the td2pLL model (right column). For the compounds PPL and GLC, separate fits might be more plausible, while for LAB and DFN the td2pLL model seems to stabilize the fit. Only concentration-wise response means are shown.

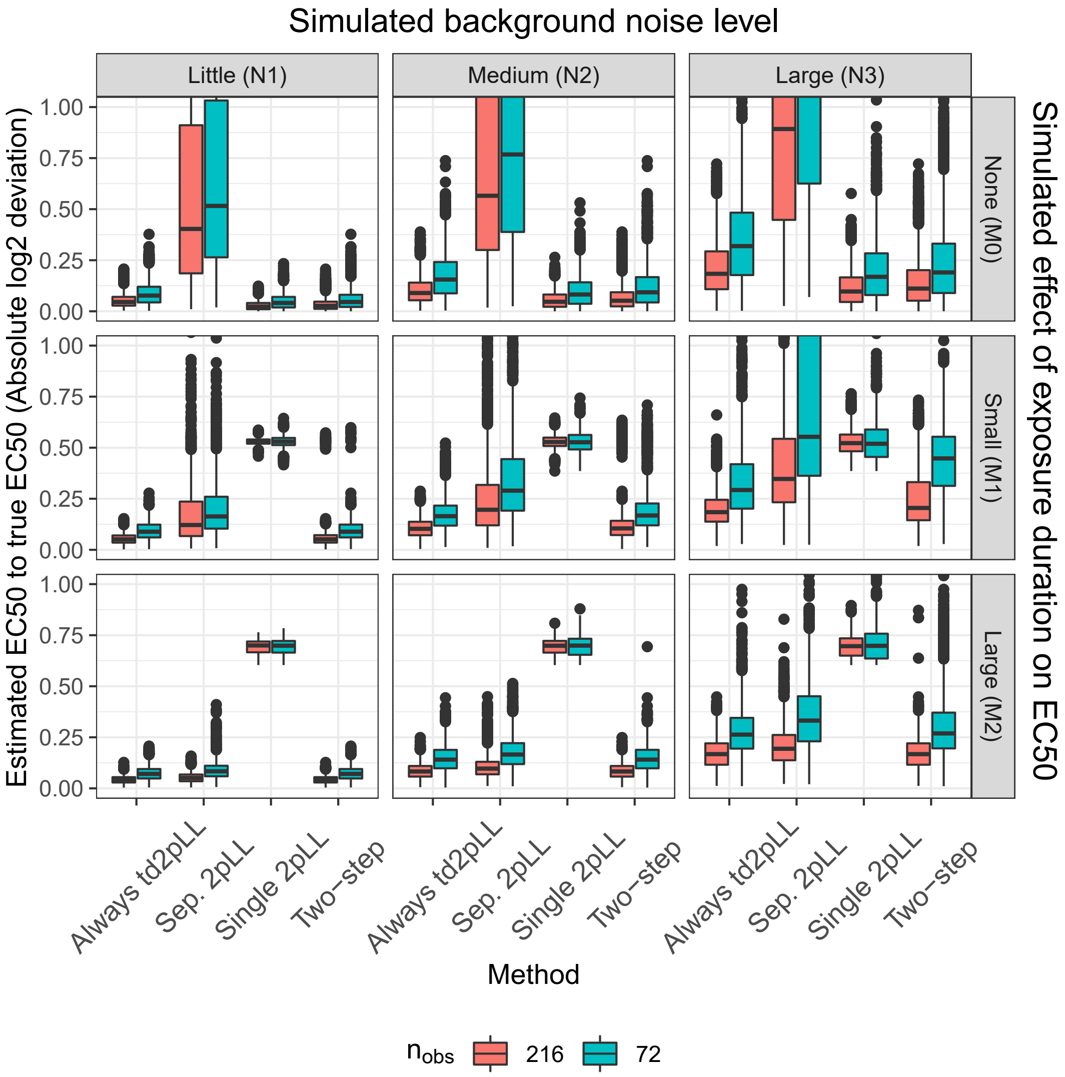


Fig. 9. The proposed methods, Two-step and always td2pLL, outperformed other methods regarding EC_{50} estimation across different simulation scenarios for exposure duration-concentration response cytotoxicity data. The three vertical panels indicate increased simulated background noise levels. The three horizontal panels correspond to an increasing effect of the exposure duration on the EC_{50}. For M_0 (upper horizontal panel), the EC_{50} is the same across all exposure durations. For M_1 (middle horizontal panel), the EC_{50} decreases moderately for increasing exposure durations, and for M_2 (lower horizontal panel) it decreases a lot for increasing exposure durations. Two-step is the proposed two-step pipeline with pre-test for exposure duration-dependency. Small values of a deviation between the estimated and the true EC_{50} are desirable.

of parameters by dividing by the degrees of freedom, i.e. the number of observations minus the number of parameters. The residual SE measures how much the model fit deviates from the data and penalizes models with more parameters. Without this penalty, approach 2 would have always fitted more closely to the data than approach 1 because of overfitting, even if the assumptions of the td2pLL model are true. If the assumptions of the td2pLL model are true or close to the truth, we expect the SE of the two approaches to be similar. The similarity of the SEs of the two approaches across the compounds can be seen in Fig. 7, in which the SEs of the approaches are close to the diagonal line. If the assumptions of the td2pLL model were markedly incorrect, the SE of approach 1 would be systematically larger and the points would systematically tend to lie below the diagonal line.

The data are not optimal to validate the td2pLL model, as there are only 3 different exposure durations and often too low maximal concentrations were used (Fig. A1).

For experiments with more exposure durations investigated, approach 2 would increasingly overfit the data, as with every added exposure duration, a new 2pLL model is fitted. By contrast, for the td2pLL model, the number of parameters remains the same, regardless of how many exposure durations are investigated. Therefore the td2pLL model would not overfit the data. In summary, the assumptions of the new td2pLL model and thus its potential benefit in target concentration estimation precision are generally supported by the analysis on the small ECR data set of [1].

For qualitative demonstration purposes, we have focused on four example compounds (PPL, DLC, LAB, DFN) where the td2pLL model appears to be either more or less adequate than separate 2pLL fits (Fig. 8). For PPL and GLC, separate 2pLL fits appear more appropriate. Especially for PPL, the assumption of a fixed steepness parameter, h, across all exposure durations seems questionable. For LAB and DFN, however, the td2pLL model fit seems to be more robust (less overfitting): For LAB, the data at the exposure duration of 7 days were very noisy, which made a separate fit for these data difficult. By fitting a td2pLL model, the concentration-response curve fit at $t = 7$ was stabilized through the (less noisy) concentration-response curves at the other exposure durations. For DFN, the concentration-response curves for exposure durations of 2 and 7 days were similar. A td2pLL model fit practically pooled and therefore stabilized the curves of these two exposure durations by fitting a large γ, i.e. a strong initial change of the

$EC_{50}(t)$ and almost no change in the $EC_{50}(t)$ for larger exposure durations.

In summary, the assumptions of the newly proposed td2pLL model for ECR data are supported by the data set of Gu et al. [1] as far as the small size of the data set allows any validations. A possible modification of the td2pLL model is a varying steepness of the concentration-response curves across exposure durations. However, this would require more parameters, which can easily lead to overfitting or computational issues. In general, the application on the real cytotoxicity data of Gu et al. [1] supports that the newly proposed td2pLL model might lead to more precise target concentration estimation for ECR data at an exposure duration of interest, by exploiting available concentration-response data of other exposure durations.

3.2. Results of the simulation study

In this section we present the results of the simulation study described in Section 2.5. In particular, the precision of the target concentration (EC_{50}) estimation for the proposed two-step pipeline that uses the ECR model, td2pLL, was compared with other approaches for estimating the EC_{50} from ECR data. The main finding was that the two-step pipeline, as well as always fitting the new td2pLL model, outperformed other approaches that ignore or overfit the potentially different EC_{50} values at different exposure durations. Fig. 9 summarizes the simulation results, comparing in each subgraph the four methods *Two-step* (new two-step pipeline described in Section 2.3), *Sep. 2pLL* (separate 2pLL models for each exposure duration), *Single 2pLL* (Single 2pLL model independent of exposure duration) and *td2pLL* (new ECR model). The subgraphs correspond to increasing noise (from left to right) and increasing exposure duration effect on the EC_{50} dependency (no, weak, strong, from top to bottom).

In general, the Two-step pipeline performed very well in all scenarios, with comparatively small deviations between estimated and true EC_{50} values. The method decides in a first step, if the EC_{50} changes across exposure durations using a nested ANOVA approach. If so, a td2pLL model is fitted that accounts for such an effect. If not, the effect of exposure duration is ignored and a single 2pLL curve is fitted for all exposure durations.

On the one hand, ignoring the exposure duration (Single 2pLL) was clearly worse, when an exposure duration-effect was present (in models M_1 and M_2, middle and bottom row subgraphs). Fitting separate 2pLL curves (Sep. 2pLL) lead to larger performance losses especially in the case without an exposure duration effect (M_0, top subgraphs). Ignoring the exposure duration and fitting a single concentration-response curve would better exploit the data structure. The performance loss was due to the much smaller effective sample size in the separate fits (Sep. 2pLL).

Furthermore, increasing the sample size (n_{obs}) naturally improved the performance (lower values for estimated to true EC_{50} deviance are observed for boxplots with red color compared to blue color). However, the beneficial effect diminished if there was only little noise (N_1) or if the method was based on severely incorrect assumptions (using Single 2pLL for scenario M_2 with strong exposure duration effect).

Lastly, always fitting a td2pLL model (without ANOVA pre-step) performed almost as well as the two-step pipeline that includes the pre-test, with only little performance loss when there was no exposure duration effect (M_0). Since using the two-step pipeline instead of always fitting the tp2pLL model can lead to a simpler model with fewer parameters (Single 2pLL), this is preferred over the (more complicated) td2pLL model. This means that in some cases, the ANOVA-based two-step pipeline avoids the fit of an unnecessarily complex model. In summary, the simulation study clearly promoted the use of the proposed two-step pipeline for target concentration estimation in ECR data.

4. Discussion

In this work we proposed a new ECR model, named td2pLL, for an improved assessment of concentration and exposure-duration dependent cytotoxicity. For cytotoxicity experiments where concentration-response data are available for different exposure durations, the proposed ECR model can improve the precision of target-concentration estimation, such as the EC_{50}, compared to fitting separate concentration-response curves for each exposure duration. This is possible because the model implicitly uses the concentration-response data from all exposure durations to calculate the target concentration at an exposure duration of interest. In addition, joint modeling allows the extrapolation of cytotoxicity responses to exposure duration-concentration combinations that were not conducted in the experiment. This is a main advantage over separately fitting concentration-response curves at each exposure duration, where cytotoxicity estimation is restricted to the exposure durations used in the experiments.

The potential benefit of the proposed ECR model over separate concentration-response curve fitting depends on the validity of the model assumptions. We presented arguments both on a theoretical and practical level, that support these model assumptions. From a theoretical perspective, the model assumptions are mechanistically motivated as it extends the well-established, one-dimensional Hill model to a two-dimensional ECR model using Haber's law. For a practical investigation, we checked the suitability of the model using a real cytotoxicity data set of 30 compounds, with measurements for multiple exposure durations and concentrations for each compound. This practical application generally supported the model assumptions, as far as the limited size of the data set allowed strong conclusions. The td2pLL model seemed to be beneficial for the data for some compounds, while for others, a more flexible model might be more suitable.

A detailed look suggested a possible extension of the td2pLL model, namely a dependency on the (currently constant) slope parameter on the exposure duration. However, such model extensions might cause problems as a more complex model is more difficult to fit, especially for typically small cytotoxicity data sets. Due to the small sample sizes and number of different exposure durations of the analyzed data, further ECR cytotoxicity data are required to assess the general benefit of the model for target-concentration estimation or if model modifications are required.

In general, exposure duration-related questions regarding the toxicological mechanism of a compound can be analyzed with the td2pLL model, such as: What is the hypothetical EC_{50} of an infinitely long exposure duration? When does expanding the exposure duration have no further influence on the toxic effect? Due to its mechanistic motivation, the model parameters can be easily interpreted to answer these questions. In fact, one model parameter is the hypothetical limit EC_{50} for an infinitely long exposure duration.

We also proposed a two-step pipeline that first performs a statistical test to decide whether a td2pLL model fit might be beneficial, or if it suffices to ignore the exposure duration and fit a regular 2pLL concentration-response curve. The two-step procedure and always fitting a td2pLL model achieved higher precision in target concentration estimation than separately fitting regular 2pLL curves per exposure duration or just fitting a single 2pLL curve for all exposure durations, as demonstrated in a simulation study based on the data of Gu et al. [1]. We recommend using the new two-step pipeline as it was successful in deciding if a td2pLL model or just a simple 2pLL model should be fitted. For practitioners, easy applicability of the proposed model and the two-step pipeline was achieved by providing the new R-package td2pLL, which is available at https://github.com/jcduda/td2pLL/.

In summary, the td2pLL model and the two-step pipeline are promising approaches to increase precision in target concentration estimation for cytotoxicity experiments in which different exposure durations are tested. The proposed methods outperformed the typical approach of separately fitting concentration-response curves for each exposure duration, as they can incorporate the data of all exposure durations for a single fit, which additionally allows cytotoxicity response estimations at concentrations and exposure durations not conducted in the experiment.

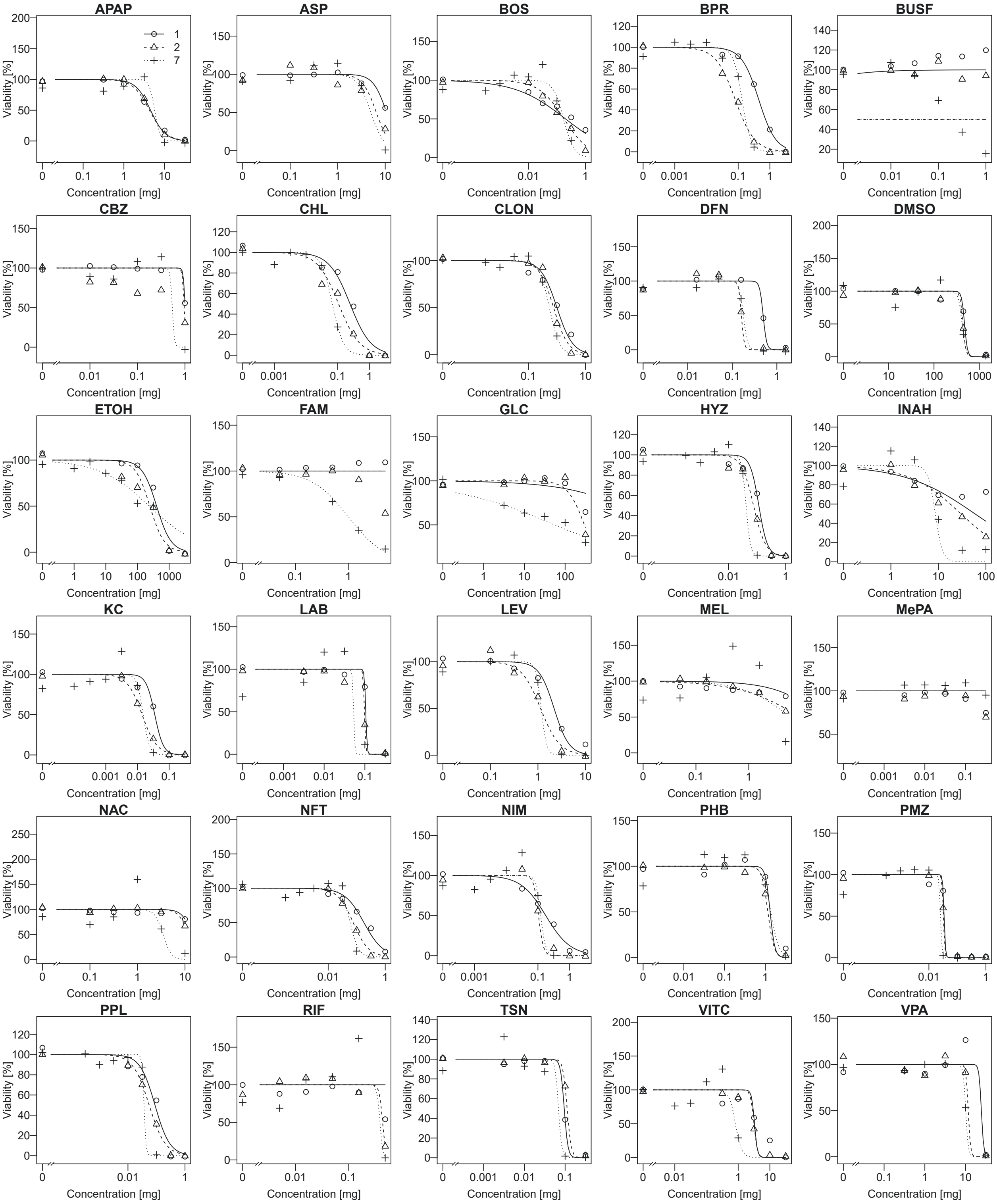


Fig. A1. Overview of separate 2pLL model fits by exposure duration (1, 2 or 7 days - legend in upper left plot) for all compounds of Gu et al. [1]. For compounds BUSF, FAM, MePA and RIF, computational problems occured as due to a high level of noise or a lack of a detectable concentration-dependent trend in viability, a flat concentration-response curve was fitted for some exposure durations. Only the concentration-wise response means are shown.

Declarations

Funding

This work has been supported (in part) by the Research Training Group ”Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project P2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116).

Data availability

The data presented in this study are openly available in Archives of Toxicology at 10.1007/s00204-018–2302-0.

Supplementary information

To reproduce all figures and the simulation results, code is available at https://github.com/jcduda/td2pLL-code.

CRediT authorship contribution statement

Julia Duda: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Jan G. Hengstler:** Validation, Resources, Writing - review & editing, Supervision, Funding acquisition. **Jörg Rahnenführer:** Conceptualization, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

Appendix A. Figures

References

[1] X. Gu, W. Albrecht, K. Edlund, F. Kappenberg, J. Rahnenführer, M. Leist, W. Moritz, P. Godoy, C. Cadenas, R. Marchan, Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes, Arch. Toxicol. 92 (12) (2018) 3505–3515.

[2] M.D. Arbo, S. Melega, R. Stöber, M. Schug, E. Rempel, J. Rahnenführer, P. Godoy, R. Reif, C. Cadenas, M. de Lourdes Bastos, Hepatotoxicity of piperazine designer drugs: up-regulation of key enzymes of cholesterol and lipid biosynthesis, Arch. Toxicol. 90 (12) (2016) 3045–3060.

[3] ICCVAM: Guidance document on using in vitro data to estimate in vivo starting doses for acute toxicity. NIH Publication NO 01-4500 (2001). Available here [accessed 29 September 2021].

[4] R. Riddell, D. Panacer, S. Wilde, R. Clothier, M. Balls, The importance of exposure period and cell type in in vitro cytotoxicity tests, Altern. Lab. Anim. 14 (2) (1986) 86–92.

[5] NICEATM: In vitro cytotoxicity test methods for estimating acute oral systemic toxicity: Background review document. NIH Publication No. 07-4518 (2006). Available here [accessed 29 September 2021].

[6] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing.https://www.R-project.org/.

[7] C. Ritz, Toward a unified approach to dose-response modeling in ecotoxicology, Environ. Toxicol. Chem. 29 (1) (2010) 220–229.

[8] C. Ritz, F. Baty, J.C. Streibig, D. Gerhard, Dose-response analysis using r, PloS One 10 (12) (2015) 0146021.

[9] F. Kappenberg, M. Grinberg, X. Jiang, A. Kopp-Schneider, J.G. Hengstler, J. Rahnenführer, Comparison of observation-based and model-based identification of alert concentrations from concentration–expression data, Bioinformatics (2021).

[10] F. Haber, Zur Geschichte des Gaskrieges, in: Fünf Vorträge aus den Jahren 1920–1923, Springer, Berlin, Heidelberg, 1924, pp. 76–92.

[11] F.J. Miller, P.M. Schlosser, D.B. Janszen, Haber's rule: a special case in a family of curves relating concentration and duration of exposure to a fixed level of response for a given endpoint, Toxicology 149 (1) (2000) 21–34.

[12] W.W. Focke, I. Van der Westhuizen, N. Musee, M.T. Loots, Kinetic interpretation of log-logistic dose-time response curves, Sci. Rep. 7 (1) (2017) 1–11.

[13] A. Schüttler, R. Altenburger, M. Ammar, M. Bader-Blukott, G. Jakobs, J. Knapp, J. Krüger, K. Reiche, G.-M. Wu, W. Busch, Map and model—moving from observation to prediction in toxicogenomics, GigaScience 8 (6) (2019) 057.

[14] A. Serra, M. Fratello, G. Del Giudice, L.A. Saarimäki, M. Paci, A. Federico, D. Greco, Tindermix: Time-dose integrated modelling of toxicogenomics data, GigaScience 9 (5) (2020) 055.

[15] F. Kappenberg, T. Brecklinghaus, W. Albrecht, J. Blum, C. van der Wurp, M. Leist, J.G. Hengstler, J. Rahnenführer, Handling deviating control values in concentration-response curves, Arch. Toxicol. 94 (11) (2020) 3787–3798.

[16] T.P. Morris, I.R. White, M.J. Crowther, Using simulation studies to evaluate statistical methods, Stat. Med. 38 (11) (2019) 2074–2102.

Article 3

scientific reports

OPEN

Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons

Julia C. Duda✉, Carolin Drenda, Hue Kästel, Jörg Rahnenführer & Franziska Kappenberg

High throughput RNA sequencing experiments are widely conducted and analyzed to identify differentially expressed genes (DEGs). The statistical models calculated for this task are often not clear to practitioners, and analyses may not be optimally tailored to the research hypothesis. Often, interaction effects (IEs) are the mathematical equivalent of the biological research question but are not considered for different reasons. We fill this gap by explaining and presenting the potential benefit of IEs in the search for DEGs using RNA-Seq data of mice that receive different diets for different time periods. Using an IE model leads to a smaller, but likely more biologically informative set of DEGs compared to a common approach that avoids the calculation of IEs.

With the rapid developments in next-generation sequencing (NGS) technology in the last decades, analyses of gene expression data have become regular in many laboratories[1]. A common goal is to identify differentially expressed genes (DEGs) that are responsible for the observable differences between, e.g., groups of individuals with different treatments or genotypes. Many software applications became available to optimally extract information from the large amounts of experimental data[2]. The mathematics behind these algorithms and models is often complicated, which can lead to suboptimal data analysis from practitioners and bioinformaticians. The interaction effect (IE) between two or more factors of interest is a methodological aspect that is often not considered in analyses where it could be beneficial. IEs are well-known in statistical modeling but are often not used in practice. Properly including and interpreting an IE in gene expression data analyses can be challenging, and the possibility of using an IE is often overlooked. An obvious reason for not using IEs in DEGs analyses might be the complexity of the statistical models and their correct computational implementation.

In the literature, there are many application examples similar to the one we will use throughout the manuscript, where an IE was likely beneficial to find interesting DEGs, but not considered. For example,[3] dealt with time-restricted feeding of mice to test whether it could prevent obesity. They used DESeq2[4] and the design included several factors such as genotype, feeding group, and time. In this setting, combining different variables to explore the interaction between e.g. time and genotype could have led to other, potentially more interesting DEGs. In another example[5] used four separate study groups to analyze the differences in heart failure in mice. They either received a standardized chow or a high-fat diet for 12 weeks, and either additionally received angiotensin II after 8 weeks or not. Here as well, analyzing the excluded interaction between diet and hormones could lead to additional interesting insights.

Examples with an IE included in the DEG analysis were provided by[6,7]. Sloley et al.[6] studied the exposure to high-frequency head impacts in mice. They use the DESeq2 package and their design contains an IE of the two factors treatment and injury. Similar methods are used in[7], in which mice were treated with acarbose at three independent study sites. Their model contains the variables treatment, sex, and the interaction between them.

In this work, we explain the use, interpretation, and potential benefit of using IEs in gene expression analysis to identify DEGs. The article equips practitioners with a less profound statistical background with the knowledge to decide if the use of an IE helps answer their research question. We therefore aim at keeping the level of mathematical complexity low, to reach a wider range of potential users. Mathematical details can be found in[8,9]. We illustrate, explain, and compare DEG analyses with and without IE using a gene expression data set from[10], where mice were fed either an unhealthy or a healthy diet for 3 to 48 weeks.

Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany. ✉email: duda@statistik.tu-dortmund.de

The article is structured as follows. We first explain the IE from different perspectives. Then we conceptually compare the use of an IE with the common approach that avoids modeling of interaction w.r.t. the resulting DEGs. The two methods are applied to the data set at hand and the differences in the results are discussed and explained in detail.

Material and methods

Data

The data set was first presented by[10], where mice were fed with two different diets over the course of 48 weeks. One diet was the high-fat or 'Western' diet (WD) and the control was a standard diet (SD). The nine analysis time points within the 48 weeks were week 3, 6, 12, 18, 24, 30, 36, 42 and 48. In total 79 samples (mice) were used. The gene expression data from 35,727 genes were measured using RNA-seq. After removing the weeks with no data from mice in one of the two groups, 64 samples from the weeks 3, 6, 30, 36, 42, and 48 were left. To focus on the explanatory aim, analyses were mostly restricted to the data of weeks 3 and 6. The sample sizes in the remaining weeks are 7, 5, 5, 7, 3, 5 for SD and 5, 5, 5, 5, 4, 8 for WD. Further pre-processing is explained in "Implementation".

Interaction effects explained

When two or more factors are of interest in an experiment, one should consider including IEs in the statistical model. Only using additive or main effects may not result in adequate modeling of the data. In Fig. 1, different effect scenarios are visualized using interaction plots for the case of two factors of interest, e.g. some group (0 = blue, 1 = red) and a compound with low and high concentration. In Fig. 1a, there is no interaction between the group and the concentration: The increase of the response from the low to the high concentration is the same for group 0 and group 1. At the same time, for a fixed concentration, the difference in the responses between group 0 and group 1 is the same. One can describe the *absence* of an IE graphically, biologically, and mathematically.

- Graphically, an additive effect or the lack of an IE results in parallel lines between the two groups.
- Biologically, the effect of the concentration does not interact with the effect of the group, because it is always the same increase in response from low to high concentration, regardless of the group.
- Mathematically, considering two factors with two levels each, a classical linear model, or equivalently an ANOVA model, with only additive effects for the two factors and normal noise is appropriate to model the data. This formalizes to

$$y_j = \mu + \alpha \cdot g_j + \beta \cdot c_j + \varepsilon_j \tag{1}$$

where j indicates the sample, g_j indicates if the jth sample is in group 0 ($g_j = 0$) or in group 1 ($g_j = 1$), and c_j indicates if the j-th sample is exposed to the low concentration ($c_j = 0$) or the high concentration ($c_j = 1$).

The mean difference in the responses for group 1 compared to group 0 is α and for increasing the concentration from low to high, the mean difference is β.

For example, if the j-th sample is in group 0 ($g_j = 0$) and exposed to the low concentration ($c_j = 0$), the expected response is $\mu + 0 \cdot \alpha + 0 \cdot \beta = \mu$. The intercept μ represents the mean response in the reference group (0) with the reference concentration (low).

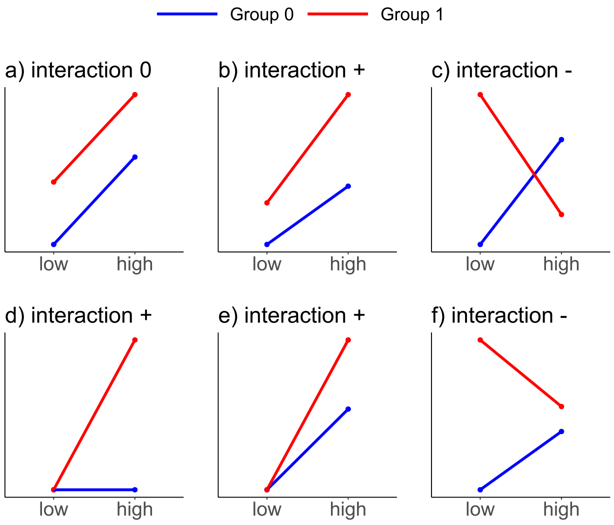


Figure 1. Schematic depiction of data scenarios without and with IE. (**a**) Group 0 (blue) and 1 (red) both have a positive effect for treatment high compared to low and a positive group effect, but no IE. (**b**) As in (**a**), but with an additional positive IE. (**c**) Negative IE between group and treatment. (**d**) No treatment effect for group 0. The treatment effect for group 1 is entirely represented by the IE. (**e**) Both groups display a positive treatment effect and there is no group effect in the treatment category low, only in high, i.e. an IE is present. (**f**) Negative IE between group and treatment, but no line crossing as in (**c**).

The contrary case, the *presence* of a clear IE with a changed direction for the concentration effect, is depicted in Figure 1c. The crossing lines mean that the effect of a concentration increase is not additive (it is not the same within both groups). Instead, the concentration effect depends on the group, i.e. there is an *interaction* with the group effect. For group 0, an increase in the concentration leads to an increase in the response, whereas for group 1, an increase in the concentration leads to a decrease in the response. The additive model (1) can not capture this interaction as the model fit would force parallel lines into the effect plot. Mathematically, a model that accounts for the interaction between group and treatment is, therefore, an extension of the model in Eq. (1) by adding the IE γ:

$$y_j = \mu + \alpha \cdot g_j + \beta \cdot c_j + \gamma \cdot g_j \cdot c_j + \varepsilon_j. \tag{2}$$

If the j-th sample is exposed to the higher concentration ($c_j = 1$) and is in group 1 ($g_j = 1$), then the mean response is $\mu + \alpha + \beta + \gamma$. The interaction term $\gamma \cdot g_j \cdot c_j$ allows the lines in the interaction plot to be non-parallel. It is important to note that an IE does not necessarily visualize as a *crossing* of lines in an interaction plot, but simply *non-parallel* lines, such as in the examples shown in Fig. 1b, d, e, and 1f. We elucidate the use of IEs when analyzing real data in the context of biological research questions in "When do interaction effects capture the research question?".

Interaction effects calculated with DESeq2

In this section, we explain the mathematical background of gene expression modeling with the popular R-package DESeq2 [4]. Details on statistical concepts presented here may not be relevant to readers who are more application-oriented and can be ignored without risking comprehension of the remaining sections. However, to understand an IE in more depth, we encourage to understand the parameters in the model formula (4).

Consider the count matrix K, where K_{ij} are the count reads of gene i for sample j, $i \in \{1, ..., n\}$, $j \in \{1, ..., m\}$. To model the count data, DESeq2 uses a generalized linear model with a negative binomial distribution $K_{ij} \sim \mathrm{NB}(\mu_{ij}, \tau_i)$ with mean μ_{ij} and gene-specific dispersion τ_i.

The mean of the observed counts $\mu_{ij} = s_j q_{ij}$ is modeled with the parameter q_{ij}, which is proportional to the expected true concentration of fragments for sample j and rescaled with a sample-specific size factor s_j. The parameter q_{ij} is modeled with a generalized linear model using the logarithmic link: $\log_2(q_{ij}) = \sum_r \beta_{ir} x_{jr}$. In a factorial design, $x_{jr} \in \{0, 1\}$ indicates if the rth explanatory variable applies to sample j, such that for the ith gene, β_{ir} is the $\log_2$FC for factor level r compared to the reference factor level.

For our application example ("Data"), the model has one factor for the diet (two values) and one factor for the week (six values). A model with the parameters for the week and diet without interaction is fitted for each gene $i, 1 \leq i \leq 35,727$. In the following, we suppress the gene index i and consider the sample (mouse) index j. The model used in DEseq2 is then

$$\log_2(q_j) = \mu + \alpha \cdot d_j + \sum_{r=2}^{6} \beta_r \cdot w_{jr}, \tag{3}$$

where μ (intercept) denotes the response at the reference (SD and week 3), and α is the WD (main) effect. The variable d_j is binary with value 0 for the SD and value 1 for the WD. The parameters $\beta_r, r \in \{2, ..., 6\}$, correspond to the week effects. The variable w_{jr} is the indicator variable for the week, i.e. $w_{j2} = 1$ only for week 6.

Now, adding an IE, the model is

$$\log_2(q_j) = \mu + \alpha \cdot d_j + \sum_{r=2}^{6} \beta_r \cdot w_{jr} + \sum_{r=2}^{6} \gamma_r \cdot d_j \cdot w_{jr}. \tag{4}$$

The parameter γ_2 denotes the IE between the factor diet and the factor week, comparing week 6 to week 3. The parameter γ_3 refers to the interaction between the diet and week, comparing week 30 to week 3, and so on. Due to the $\log_2$ transformation for the sample concentration q_j, the parameters must all be interpreted accordingly. For example, an IE of $\gamma_2 = 3$ means that the difference between the diet effect in week 3 and the diet effect in week 6 is $2^3 = 8$, or has a FC of 8.

When do interaction effects capture the research question?

In RNA-Seq experiments, often the case of two factors, e.g. treatment and genotype, are analyzed, and it is of interest whether the effect of the treatment differs between the genotypes (in certain genes). The research question might be formulated as: Does the genotype affect the treatment effect? IEs capture such a research question and they should therefore be considered for the analysis.

In our application example, the two factors are diet and week, where diet is either a WD or a SD and week indicates the feeding duration. In this dataset measurements for different time points are available, and we focus on the two shortest durations, 3 weeks and 6 weeks, to explain the IE concept. The 3-week time point can be considered the reference level of the factor week. The research goal is to identify genes where activation/deactivation from weeks 3 to 6 induced by the WD is different compared to the SD. Mathematically, this research question translates into identifying genes with an IE between diet and week. Consequently, the use of a model that includes an IE should be considered.

How do interaction effects capture the research question?

To explain how IEs capture the research question, we visualize the benefit of adding IEs to a linear model, using our example dataset. In Fig. 2, for the mice groups, for each combination of diet type and week, expression values and fitted means are plotted, exemplary for one selected gene. Once no IEs are included in the model (Fig. 2, left), and once IEs are included (Fig. 2, right).

Without IEs, the estimated effect differences between the diets, represented by arrows, are mathematically forced to be the same across all weeks (vertical lines have the same length).

Consequently, in week 3, the effect is markedly overestimated, as the arrow between SD and WD is larger than the pure difference in group means. In contrast, if an IE is used (Fig. 2, right), then the group means estimated by the model capture well that the diet effect varies across weeks. The mathematical formulas of the estimated effects represented by the arrows are explained in "Interaction effects calculated with DESeq2".

Comparison of methods for estimating interaction effects

In this section, we compare the results obtained by fitting an interaction model between two factors (called Method II in the following) with a far more popular alternative, which we call Method I. The alternative approach avoids the direct modeling of an IE between two factors as follows: The data are split with respect to the second factor (e.g. week) into two groups G_0 and G_1. Then for group G_0 and G_1 separately, a model comparing the groups with respect to the first factor (e.g. diet) is fitted. Finally, it is analyzed, if for one group, typically the reference group G_0, no significant effect is observed, and for the other group G_1, there is a significant effect present.

The differences between the two approaches are illustrated and discussed on the mouse dataset, where for Method I the groups G_0 and G_1 are defined by week 3 (as reference) and week 6 (or larger week numbers, respectively). The models per week contain only one factor (diet) with two levels, SD and WD. Since separate models are fitted per week, the model-wise diet effect is allowed to vary across weeks.

When interpreting the results of the differential expression analysis, a consideration of both *statistical significance* and *biological relevance* is necessary: A *p*-value smaller than the significance level, which constitutes a statistically significant result, does not necessarily mean that the mean effect level, given here by the $\log_2$-Fold Change ($\log_2$FC), is of relevant size. On the other hand, a mean effect level larger than a pre-specified threshold, motivated by the biological context, does not always correspond to small *p*-values [11]. Thus, to interpret a gene to be a differentially expressed gene (DEG), we always require two conditions to be fulfilled: The (FDR-adjusted) *p*-value is smaller than a significance level, and the $\log_2$FC is larger than a pre-specified threshold.

For the mouse dataset and the separate models (Method I), only those genes that show a diet effect (both significant and relevant) in week 6, but not in the reference week 3, are considered DEGs. The motivation is that interesting genes show no effect at the reference time point, where the diet had too little time to cause a differential effect, but later (at 6 weeks) the diet causes such a difference. For the interaction model (Method II), not two models but only a single model is fitted. To detect DEGs, one simply checks if the estimated IE is both significant and relevant.

- *Method I* (Separate): Separately for each week: Fit a one-factor model (two-group comparison, see equation (1)).

  A gene is DEG if the diet effect is both significant and relevant in week 6, but not both in week 3.
- *Method II* (Interaction): Fit a two-factor model between week and diet (including week, diet, and interaction), see equation (2).

  A gene is DEG if the IE is both significant and relevant.

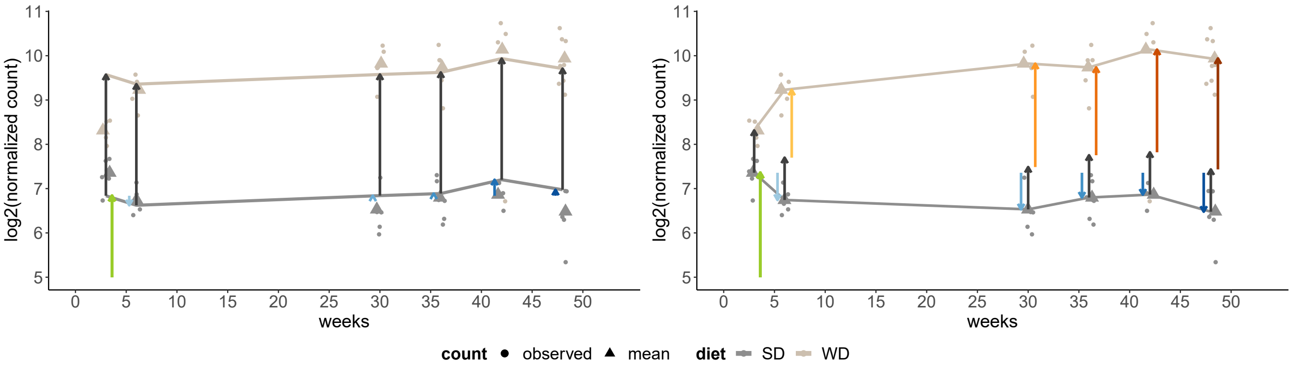


Figure 2. Visualization of the fitted model without IE (left) and with IE (right) for the mice dataset, for the gene identifier ENSMUSG00000069170 (Adgrv1). The arrows represent the estimated $\log_2$FCs according to Eq. (3) for the left fit, and Eq. (4) for the right fit. For both fits, μ (green arrow) is the expected mean gene expression level for the reference values three weeks and SD, and α (vertical dark grey arrows) is the estimated FC between SD and WD at each week. Further, both models include the week effects β_r (blue arrows). The right model additionally includes interaction effects (yellow, orange, and red arrows) that correspond to γ_r in formula (4).

To visualize the differences between the decision outcomes (gene is DEG or not) of Method I and II, Fig. 3 displays 7 cases using simulated data scenarios. The data are generated with constant residual variance, so that the decision is not influenced by differing variance values, but only by the estimated effect (arrow lengths).

- Case 1: Within both weeks, the estimated diet effect is not relevant (dotted green effect arrow). There is hence is no DEG by Method I. Since the effects are of similar size, the IE estimated by Method II (pink arrow) is not significant, and neither Method II classifies the gene as DEG.
- Case 2: In week 3, the effect is not relevant, in week 6 it is both significant and relevant. This leads to a significant IE for Method II. Therefore, both Method I and Method II classify the gene as DEG.
- Case 3: The diet effect is significant in both weeks. Since it is significant in week 3, Method I does not classify the gene as DEG. However, the diet effect in the second week is much larger, such that the IE is significant, and Method II classifies the gene as DEG.
- Case 4: Similar to case 3, but the effect direction of the diet effect changes: In the first week, there is a positive effect, and in the second week a negative effect. Again, only Method II classifies the gene as DEG.
- Case 5: In week 3, the diet effect is just below the significance level, whereas in week 6 it is just above the significance level. Therefore, Method I labels the gene as DEG. For Method II, the IE is not significant as the diet effect does not differ much between the weeks. Method II does not label the gene as DEG.
- Case 6: Similar to case 4, but the effect in week 3 is not significant. Now both methods classify the gene as DEG.
- Case 7: The direction of the diet effect changes. It is positive in week 3 and negative in week 6. Within each week, the effect size is not significant, therefore Method I classifies the gene as not DEG. The overall change in the effect represented by the IE is significant. Therefore, Method II labels this gene as DEG.

Implementation

For all calculations, R[12], version 4.2.2, and the packages `DESeq2`[4], version 1.38.1, and `topGO`[13], version 2.50.0, were used for determining DEGs and performing gene ontology enrichment analyses (GO EA), respectively. The entire code is shared on *GitHub* (https://github.com/jcduda/gene_expression_interaction). We specify the models of Method I and II in `DESeq2` using

- Method I: `DESeqDataSet(gse, design =~diet)`
- Method II: `DESeqDataSet(gse, design =~diet + weeks + diet:weeks)`

In the example, the code for Method I is applied twice for separate weeks, i.e. for two different data sets '`gse`', while the code for Method II is applied only once. Note that a model based on `~diet + weeks` results in the same parameter values for each week, making it unsuitable for comparison with Method I and Method II, see Fig. 2.

One notable preprocessing step was the filtering. Removing only genes with less than ten counts over all samples resulted in a peak of the estimated diet effect at 0.206 (Supplementary Fig. 1). However, removing genes with more than 50% of samples with 0 counts leads to reasonably estimated effects without artifactual spikes in the histogram (Supplementary Fig. 2). Further, we shrunk the estimated effects using approximate posterior estimation with the *lfcShrink* function[14]. Effects that are non-zero only due to noise are shrunk to zero, while large, reliable effects are not affected.

Results

We compare Method I (separate) and Method II (interaction) for the mouse dataset, w.r.t. classification of genes as DEG or not DEG, as described in "Comparison of methods for estimating interaction effects". In the following list, we define the terms significant, relevant, and DEG in the context of the example study.

For Method I we call a gene

- significant, if false discovery rate (FDR) adjusted p-value < 0.05 (for a specific week X)
- relevant, if absolute $\log_2$FC $> \log_2(1.5)$ (for a specific week X)
- DEG for week X, if it is significant and relevant for week X
- DEG, if it is not DEG for week 3, but DEG for week 6

For Method II we call a gene

- significant, if FDR adjusted p-value < 0.05 (for the IE)
- relevant, if absolute $\log_2$FC $> \log_2(1.5)$ (for the IE)
- DEG, if it is significant and relevant (for the IE)

For Method I, up-regulated DEGs for week X have a positive diet effect in week X. For Method II, up-regulated DEGs have a positive IE. Down-regulated DEGs are defined accordingly.

Comparison of genes selected by Method I and Method II

We expect a relevant number of DEGs, since a biological effect of the diet (WD vs. SD) is reported by[10]. Table 1 shows the number of DEGs in week 3 and DEGs in week 6, according to Method I (simple comparison per week). There are more DEGs after 6 weeks of feeding compared to 3 weeks, both for up- or down-regulation.

For up-regulated genes, 104 genes are DEGs only for week 3, 81 genes that are DEGs in both weeks, and 1,622 genes that are DEGs only in week 6. Hence, for Method I, regarding up-regulation, one would focus on the 1622 DEGs that are only identified for week 6 and not for week 3.

Table 2 presents a main finding of our study, a comparison of DEGs identified with Method I and Method II. One can see that Method I (separate) identifies more DEGs than Method II (interaction). However, the DEGs identified by Method II are not all contained in the DEGs identified by Method I. There are almost 200 genes only identified by Method II, both for up-regulation and for down-regulation.

Characterization of genes that are DEG only for Method I or only for Method II

To understand the benefits of the two methods, we characterize the genes that are only identified by one of the two approaches, respectively. After a mathematical characterization, we also investigate biological differences.

An insightful example is gene Sirt7 in Fig. 5, which is a typical case for being DEG by Method II, but not by Method I. From week 3 to week 6, there is an interaction between the factor week and diet (crossing of grey lines). The IE (large yellow arrow) is significant and relevant, making this gene DEG for Method II. However, for Method I the $\log_2$FC of the diet effect in week 6 is not large enough to pass the threshold of $\log_2(1.5)$. Hence, Sirt7 is not identified as DEG by Method I, even though an important underlying diet effect dependent on the time seems reasonable. Such genes are overlooked by the popular Method I.

To better understand the differences between the two approaches, Fig. 6 shows regions of genes classified as DEG by both, none, or only one of the two methods, dependent on the main effect (diet) and the IE, as obtained by the interaction model (2) used by Method II.

Each dot represents a single gene. If there is no interaction (cf. Fig. 1a), the estimated IE is (close to) 0, such that the x- and y-value are identical and the gene is on the diagonal. For better illustration, the estimated effects are not shrunk and the decision rule depends on the $\log_2$FC threshold only. In practice, $\log_2$FC estimates should be subject to shrinkage and the classification into a DEG depends on both, $\log_2$FC and adjusted p-value (Supplementary Fig. 3 in the Appendix).

The genes can be divided into four groups according to the DEG classification of Method I and Method II. The numbers 1–7 assigned to regions match the simulated cases in Fig. 3 and a real gene expression pattern of a representative gene shown in Fig. 4. In the following, the gene expression patterns corresponding to the colored regions in Fig. 6 are explained.

- *Orange: not DEG for both methods.* Genes closer to the diagonal than $\log_2(1.5)$, such that the IE is below this threshold and the gene is not DEG for Method II. Further, genes with absolute main effect above $\log_2(1.5)$ are DEG for week 3 and thus not DEG for Method I.
- *Green: DEG only for Method I.* Genes with absolute main effect and IE less than $\log_2(1.5)$, but overall effect in week 6 greater than $\log_2(1.5)$. These genes are not DEG in week 3 by being slightly below the threshold but are DEG in week 6 by being slightly above the threshold. Hence, they are DEG for Method I, but the IE is small and the gene is not DEG for Method II.
- *Purple: DEG for Method I and II.* Genes with an estimated main effect (for week 3) below the $\log_2$FC boundaries, but the sum of main and IE (diet effect for week 6) is outside these boundaries. Hence, these genes are DEG for Method I. For Method II, they are DEG since the IE is large enough (points far from the diagonal line).
- *Blue: DEG only for Method II.* Genes that are not DEG for Method I since they are either DEG in week 3 (main effect outside $\pm\log_2(1.5)$) or have a main effect inside $\pm\log_2(1.5)$ (as gene 7) but are not DEG in week 6, since the corresponding effect (main plus IE) is also within $\pm\log_2(1.5)$).

We further looked at differences concerning the biological conclusions of the found DEGs. First, a qualitative, small literature research on the top 10 (lowest adj. p-value) upregulated DEGs found only by Method I or only by Method II, respectively, suggests that both methods find genes that are reasonably associated with liver disease induced by a fatty diet (Table 4; Supplementary Table 1). On a broader scale, a GO EA was performed on the DEGs found by Method I, Method II, and the combination of both DEG sets (Table 3; Supplementary Table 2). Despite the smaller number of DEGs identified by Method II, the biological interpretation based on the processes identified by GO EA is very similar and plausibly covers immune activation related to fatty liver disease. This suggests that the DEGs found by Method II are more specific in the sense that they include fewer non-relevant genes while yielding similar GO EA results.

Discussion

Using an IE model with 2 factors (Method II) instead of two separate models with one factor each (Method I) clearly changes the set of DEGs found in a gene expression analysis. The set of DEGs found with Method II is usually smaller. A theoretical reason for this is that statistical inference that aims at detecting IEs is less powerful in the sense that the sample size must be four times larger to have the same power for detecting an IE than to detect a main effect[15,16], p. 100f.

Further, a gene that just passed the thresholds for being DEG for the reference group, but just not for the other group, is DEG for Method I but usually not for Method II, and it is not a good candidate for a biologically meaningful statement. The resulting DEGs for Method II are smaller in number, but lead to equally reasonable biological findings based on enrichment analyses. A limitation of Method II is that a single model with two main factors and an IE can be more difficult to interpret correctly than two models with one factor each and no IE. Quantifying if the smaller set of DEGs found by Method II contains less irrelevant genes is difficult for several

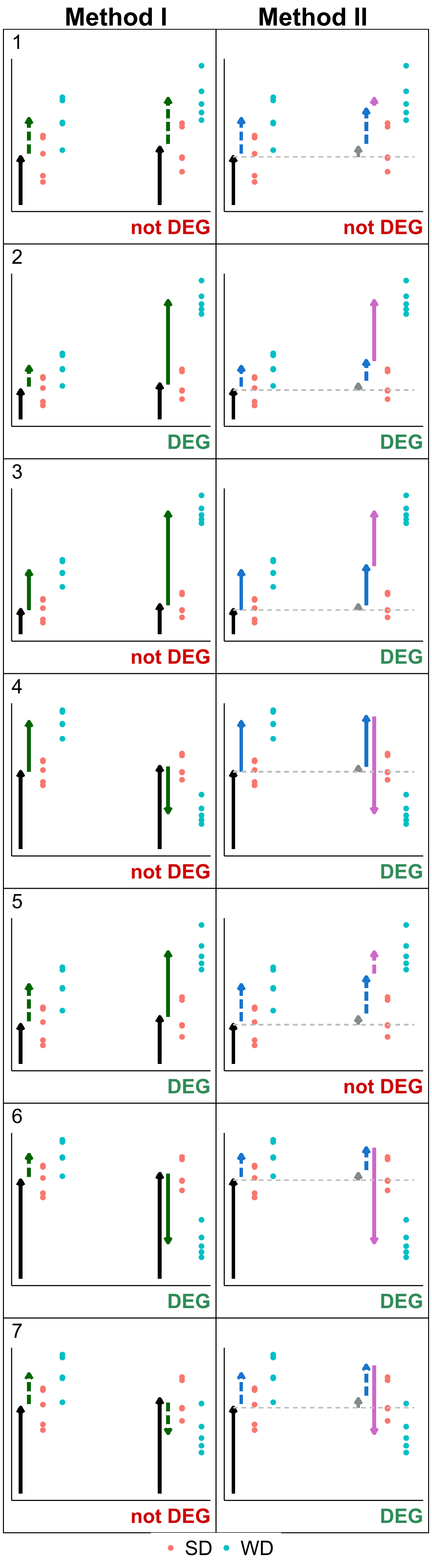


Figure 3. Visualization of seven example scenarios with different main effects and IEs, leading to different decisions for Method I (left column) and Method II (right column). Dots represent data points (blue: SD, red: WD; left: 3 weeks, right: 6 weeks), arrows represent effects (black: reference mean, green: main effect of diet, purple: IE). Dotted arrows indicate non-relevance (absolute effect size below threshold), solid arrows represent relevant effects. Dotted arrows are only shown for the main effects of IEs. The label 'DEG' below a scenario indicates if the respective method classifies a gene as DEG (green) or not DEG (red).

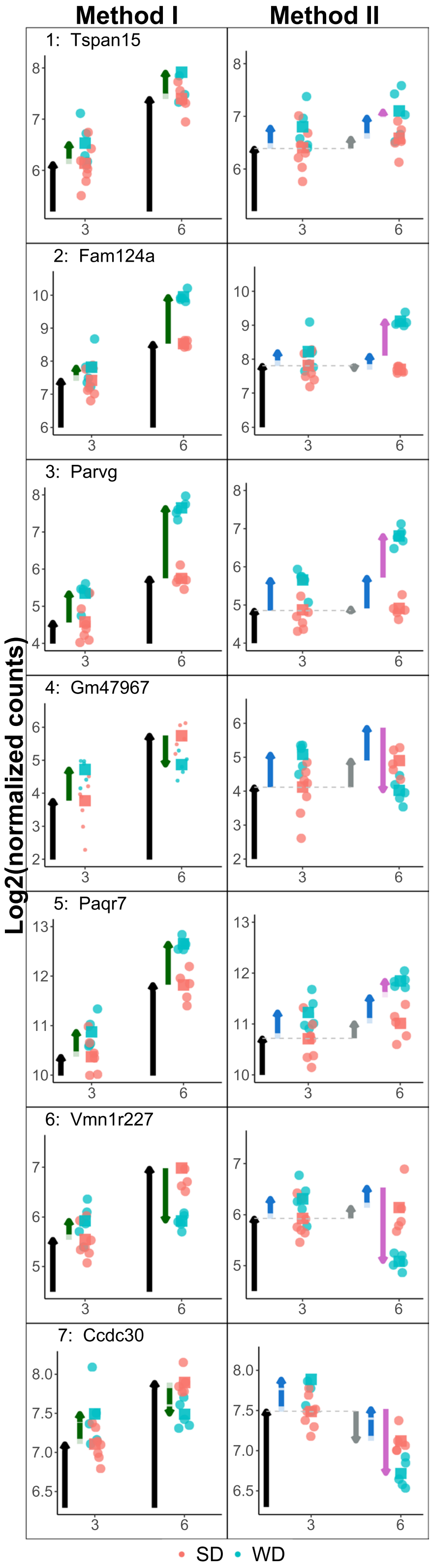


Figure 4. Example genes that are, according to DEG decision cases 1–7, not always classified in the same way by Method I (left) and II (right). Note that the original data are the same per gene (row), but due to the differences between Method I and II, background normalizations yield slightly different data for each gene. For normalization, DESeq estimates the library sizes as the median of the ratios of observed counts[9]. See caption of Figure 3 for an explanation of the arrows.

	Week 3 only	Overlap	Week 6 only
Up	104	81	1,622
Down	81	93	726

Table 1. Overview of DEGs for Method I, comparison of SD and WD.

	Method I only	Overlap	Method II only
Up-regulated	914	695	167
Down-regulated	540	177	186

Table 2. Comparison of DEGs identified with Method I and Method II. Note that 914 + 695 = 1609 does not equal 1622 in Table 1, because here we do not include genes that are downregulated in week 3, as otherwise they would not be DEG by Method I.

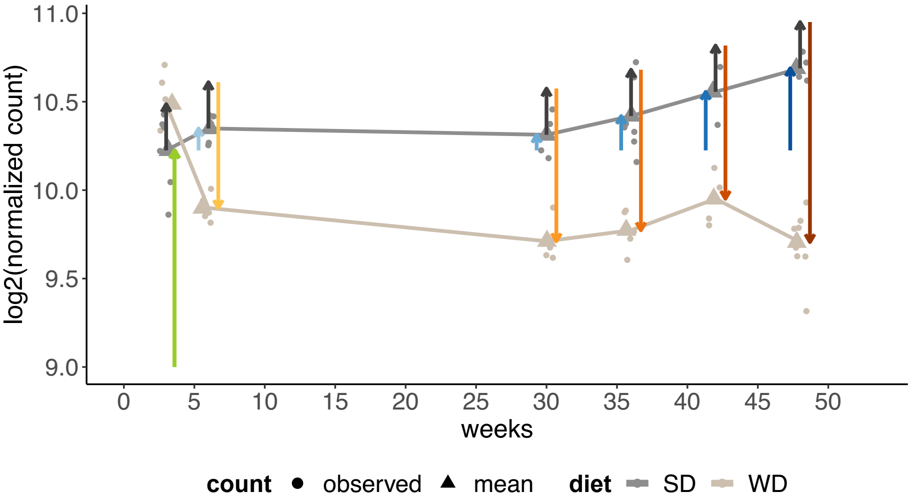


Figure 5. Expression pattern for the gene Sirt7, which is for the comparison week 3 vs. week 6 DEG for Method II (interaction), but not by Method I (separate), since the effect size is too low for week 6. See caption of Fig. 2 for detailed explanation of the arrows.

reasons. First, a literature search to determine if a gene is not reported within the context of liver disease is fruitless. Due to false positive results and extensive research in this area, almost any gene can be found as associated. Second, the data set at hand does not have a clean reference, because mice were already fed with HFD for three weeks in the reference group, instead of being fed for zero weeks. However, within the limits of this study, the conceptual reasoning and analyses of GO enrichment analyses suggest that gene sets identified by Method II are smaller but likely contain fewer irrelevant genes.

Conclusion

An IE might often be an adequate translation of a biological research question into a statistical concept. However, this relationship might remain unnoticed due to a lack of expertise or reluctance to deviate from routines. In this work, we offer an extensive explanation of IEs and why they might be scientifically relevant in the context of detecting differentially expressed genes (DEGs) in gene expression analysis.

We compare the IE-based approach (Method II) with a popular alternative approach (Method I) that avoids the calculation of IEs. While Method I detects more DEGs, many of them might not be scientifically relevant, whereas the smaller set of DEGs found with Method II can be interpreted as more specific by having fewer irrelevant genes. We encourage researchers to clarify for each project if an IE is the accurate mathematical representation of the formulated research question and to use this concept when appropriate. Further, if the research goal is to identify a smaller gene set containing less irrelevant genes (less false positives), we encourage

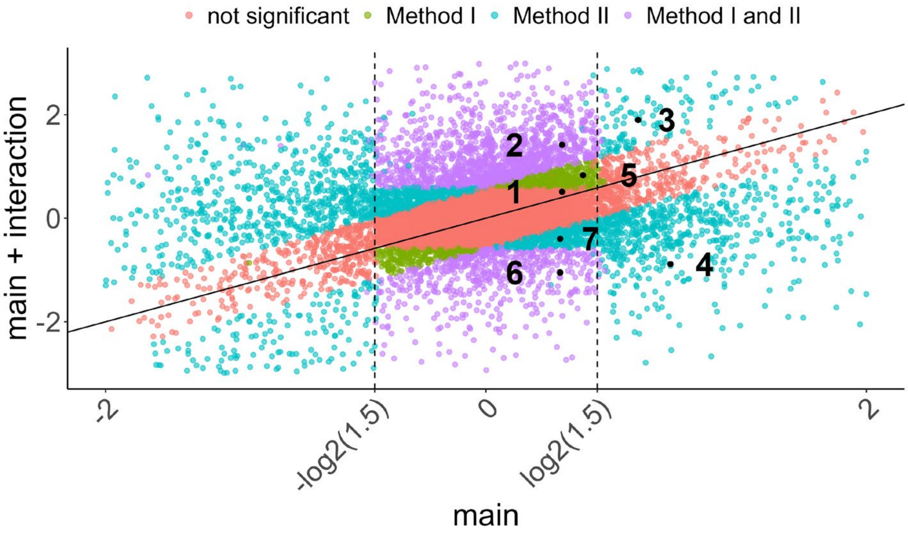


Figure 6. Characterization of regions of genes that are identified as DEG only by Method I or by Method II, or by both or none of the methods. The *x*-axis shows the estimated main effect (diet), i.e. the estimated $\log_2$FC from a SD to WD in the reference week 3, and on the *y*-axis the sum of this main effect and the IE, i.e. the overall effect between the two diets in week 6 in the interaction model, is plotted.

	Method I	Method II	Method I or II
1	Immune system process (2.44×10^{-29})	Immune system process (3.33×10^{-28})	Immune system process (2.27×10^{-29})
2	Immune response (2.44×10^{-29})	Immune response (3.33×10^{-28})	Immune response (2.27×10^{-29})
3	Defense response (2.44×10^{-29})	Cell activation (3.33×10^{-28})	Defense response (2.27×10^{-29})
4	Pos. reg. of immune system process (2.44×10^{-29})	Response to external stimulus (5×10^{-28})	Regulation of immune system process (2.27×10^{-29})
5	Regulation of immune system process (2.44×10^{-29})	Defense response (6×10^{-28})	Pos. reg. of immune system process (2.27×10^{-29})
6	Response to other organism (2.44×10^{-29})	Response to stimulus (1.65×10^{-27})	Response to external stimulus (2.27×10^{-29})
7	Response to external biotic stimulus (2.44×10^{-29})	Leukocyte activation (2.57×10^{-27})	Response to biotic stimulus (2.27×10^{-29})
8	Response to biotic stimulus (2.44×10^{-29})	Regulation of immune system process (1.2×10^{-25})	Response to other organism (2.27×10^{-29})
9	Response to external stimulus (2.44×10^{-29})	Response to external biotic stimulus (2.27×10^{-25})	Response to external biotic stimulus (2.27×10^{-29})
10	Defense response to other organism (2.44×10^{-29})	Response to other organism (2.27×10^{-25})	Defense response to other organism (2.27×10^{-29})
11	Innate immune response (2.44×10^{-29})	Response to biotic stimulus (2.27×10^{-25})	Biol. proc. involved in interspecies interaction btw organisms (2.27×10^{-29})
12	Cell activation (2.44×10^{-29})	Pos. reg. of immune system process (2.92×10^{-25})	Cell activation (2.27×10^{-29})
13	Biol. proc. involved in interspecies interaction btw organisms (2.44×10^{-29})	Pos. regulation of multicellular organismal process (4.31×10^{-25})	Pos. regulation of multicellular organismal process (2.27×10^{-29})
14	Inflammatory response (2.44×10^{-29})	Biol. proc. involved in interspecies interaction btw organisms (8.57×10^{-25})	Inflammatory response (2.27×10^{-29})
15	Pos. reg. of response to external biotic stimulus (2.44×10^{-29})	Pos. reg. of response to stimulus (2.93×10^{-22})	Innate immune response (2.27×10^{-29})

Table 3. Top 15 most significant GO groups found based on upregulated DEGs by Method I, Method II and combining the genes found by Method I and Method II. FDR-adjusted p-values are in parentheses.

to use Method II. However, if the research goal is rather exploratory and more false positives are acceptable, we suggest to use Method I.

Data availability

The analyzed data sets are publicly available at the SRA database with reference number PRJNA953810.

Code availability

The code is available on *GitHub* (https://github.com/jcduda/gene_expression_interaction).

Appendix

See Table 4.

DEG only in	Gene	Log2FC	FDR-adj. p	Literature
Method I	Acnat2 (ENSMUSG00000060317)	2.02	< 0.01	Considered a candidate for specific metabolic processes within the Type I acyl-CoA thioesterase/acyltransferase gene family[17]
	Tlr12 (ENSMUSG00000062545)	1.45	< 0.01	Signaling in chronic liver diseases via complex immune responses mediating hepatocyte[18]
	Fgf21 (ENSMUSG00000030827)	2.39	< 0.01	Associated with development and progression of NAFLD[19]
	Tgfbi (ENSMUSG00000035493)	0.79	< 0.01	Overexpression in mice resulted in an increased incidence of spontaneous tumors and N,N-diethyl-nitrosamine (DEN)-induced liver tumor nodules[20]
	Ehhadh (ENSMUSG00000022853)	0.90	< 0.01	Associated with development of fatty liver disease in dairy cows[21]
	Hpgds (ENSMUSG00000029919)	2.41	< 0.01	Overexpression associated with adipogenesis and increased insulin sensitivity[22]
	Slc17a4 (ENSMUSG00000021336)	1.14	< 0.01	An intestinal organic anion exporter expressed predominantly in the pancreas, liver, colon, and small intestine[23]
	Lgmn (ENSMUSG00000021190)	0.76	< 0.01	Elevated expression of LGMN is reported in the tumor cells of liver[24]
	Slc7a8 (ENSMUSG00000022180)	1.26	< 0.01	Slc7a8 deletion is protective against diet-induced obesity[25]
	Pgm3 (ENSMUSG00000056131)	1.17	< 0.01	Up-regulated in livers of high fat diet fed mice[26]
Method II	Mmp12 (ENSMUSG00000049723)	4.16	< 0.01	Mmp12 is a matrix metalloproteinase and associated with liver disease and inflammation[27]
	Cyp2c38 (ENSMUSG00000032808)	2.47	< 0.01	Cyp2c family up-regulated in NAFLD mouse model[28]
	Itgam (ENSMUSG00000030786)	2.92	< 0.01	Increase in Itgam (aka Cd11b) in liver X receptors knockout mice[29]
	Adgrg2 (ENSMUSG00000031298)	2.76	< 0.01	Found upregulated in cholestasis liver tissue compared to mildly damaged liver tissue[30]
	Nap1l1 (ENSMUSG00000058799)	0.71	< 0.01	Tumor promoter in hepatocellular carcinoma[31]
	Gstm3 (ENSMUSG00000004038)	2.40	< 0.01	Associated with acute-on-chronic hepatitis B liver failure[32]
	Myo1c (ENSMUSG00000017774)	0.61	< 0.01	Upregulation associated with human chronic liver disease[33]
	Rac2 (ENSMUSG00000025003)	1.89	< 0.01	Associated with NAFLD[34]
	Cyp2c39 (ENSMUSG00000025003)	2.65	< 0.01	Cyp2c family up-regulated in NAFLD mouse model[28]
	Gm3776 (ENSMUSG00000111709)	3.11	< 0.01	Gm3776, or glutathione *S*-transferase (GST) alpha 13, belongs to GST genes that are associated with liver disease[35]

Table 4. Top 10 most significant up-regulated genes found only by Method I or Method II, respectively.

Received: 14 March 2023; Accepted: 8 November 2023
Published online: 27 November 2023

References

1. Murray, D., Doran, P., MacMathuna, P. & Moss, A. C. In silico gene expression analysis—An overview. *Mol. Cancer* **6**, 1–10 (2007).
2. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-seq differential expression analysis: An extended review and a software tool. *PloS one* **12**, e0190152 (2017).
3. Chaix, A., Lin, T., Le, H. D., Chang, M. W. & Panda, S. Time-restricted feeding prevents obesity and metabolic syndrome in mice lacking a circadian clock. *Cell Metab.* **29**, 303–319 (2019).
4. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).
5. Withaar, C. *et al.* The effects of liraglutide and dapagliflozin on cardiac function and structure in a multi-hit mouse model of heart failure with preserved ejection fraction. *Cardiovasc. Res.* **117**, 2108–2124 (2021).
6. Sloley, S. S. *et al.* High-frequency head impact causes chronic synaptic adaptation and long-term cognitive impairment in mice. *Nat. Commun.* **12**, 1–20 (2021).
7. Smith, B. J. *et al.* Changes in the gut microbiome and fermentation products concurrent with enhanced longevity in acarbose-treated mice. *BMC Microbiol.* **19**, 1–16 (2019).
8. Turner, J. R. & Thayer, J. Introduction to Analysis of Variance: Design, Analyis & Interpretation: Design, Analysis & Interpretation. (Sage, 2001).
9. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nat. Prec.* 1–1 (2010).
10. Ghallab, A. *et al.* Spatio-temporal multiscale analysis of western diet-fed mice reveals a translationally relevant sequence of events during NAFLD progression. *Cells* **10**, 2516 (2021).
11. Hothorn, L. A. Statistics in Toxicology Using R. (CRC Press, 2015).
12. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2022).
13. Alexa, A. & Rahnenfuhrer, J.topGO: Enrichment Analysis for Gene Ontology. *R Package Version 2.50.0.* (2022).
14. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).
15. Leon, A. C. & Heo, M. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Comput. Stat. Data Anal.* **53**, 603–608 (2009).
16. Fleiss, J. L. Design and Analysis of Clinical Experiments. (Wiley, 2011).
17. Reilly, S.-J. *et al.* A peroxisomal acyltransferase in mouse identifies a novel pathway for taurine conjugation of fatty acids. *FASEB J.* **21**, 99–107 (2007).
18. Kiziltas, S. Toll-like receptors in pathophysiology of liver diseases. *World J. Hepatol.* **8**, 1354 (2016).
19. Tucker, B., Li, H., Long, X., Rye, K.-A. & Ong, K. L. Fibroblast growth factor 21 in non-alcoholic fatty liver disease. *Metabolism* **101**, 153994 (2019).
20. Han, B. *et al.* The role of tgfbi (βig-h3) in gastrointestinal tract tumorigenesis. *Mol. Cancer* **14**, 1–12 (2015).
21. Le-Tian, Z. *et al.* Protein acetylation in mitochondria plays critical functions in the pathogenesis of fatty liver disease. *BMC Genomics* **21**, 1–17 (2020).
22. Fujitani, Y. *et al.* Pronounced adipogenesis and increased insulin sensitivity caused by overproduction of prostaglandin d2in vivo. *FEBS J.* **277**, 1410–1419 (2010).

23. Togawa, N., Miyaji, T., Izawa, S., Omote, H. & Moriyama, Y. A Na+-phosphate cotransporter homologue (slc17a4 protein) is an intestinal organic anion exporter. *Am. J. Physiol.-Cell Physiol.* **302**, C1652–C1660 (2012).
24. Reddy, B. D., Beeraka, N. M., Chitturi, C. & Madhunapantula, S. V. An overview of targeting legumain for inhibiting cancers. *Curr. Pharmaceut. Des.* **27**, 3337–3348 (2021).
25. Pitere, R. R., van Heerden, M. B., Pepper, M. S. & Ambele, M. A. Slc7a8 deletion is protective against diet-induced obesity and attenuates lipid accumulation in multiple organs. *Biology* **11**, 311 (2022).
26. Jang, J.-H. *et al.* Klhl3 deficiency in mice ameliorates obesity, insulin resistance, and nonalcoholic fatty liver disease by regulating energy expenditure. *Exp. Mol. Med.* **54**, 1250–1261 (2022).
27. Naim, A., Pan, Q. & Baig, M. S. Matrix metalloproteinases (MMPS) in liver diseases. *J. Clin. Exp. Hepatol.* **7**, 367–372 (2017).
28. Xiang, L. *et al.* Comparison of hepatic gene expression profiles between three mouse models of nonalcoholic fatty liver disease. *Genes Dis.* **9**, 201–215 (2022).
29. Endo-Umeda, K. *et al.* Liver x receptors regulate hepatic f4/80+ cd11b+ Kupffer cells/macrophages and innate immune responses in mice. *Sci. Rep.* **8**, 1–14 (2018).
30. Liu, X., Taylor, S. A., Celaj, S., Levitsky, J. & Green, R. M. Expression of unfolded protein response genes in post-transplantation liver biopsies. *BMC Gastroenterol.* **22**, 380 (2022).
31. Zhang, Y.-W. *et al.* Nap1l1 functions as a tumor promoter via recruiting hepatoma-derived growth factor/c-jun signal in hepatocellular carcinoma. *Front. Cell Dev. Biol.* **9**, 659680 (2021).
32. Sun, F.-K. *et al.* High promoter methylation levels of glutathione-S-transferase m3 predict poor prognosis of acute-on-chronic hepatitis b liver failure. *Hepatol. Res.* **47**, 566–573 (2017).
33. Arif, E. *et al.* Targeting myosin 1c inhibits murine hepatic fibrogenesis. *Am. J. Physiol.-Gastrointest. Liver Physiol.* **320**, G1044–G1053 (2021).
34. Zhu, J., Min, N., Gong, W., Chen, Y. & Li, X. Identification of hub genes and biological mechanisms associated with non-alcoholic fatty liver disease and triple-negative breast cancer. *Life* **13**, 998 (2023).
35. Prysyazhnyuk, V. *et al.* Glutathione-S-transferases genes-promising predictors of hepatic dysfunction. *World J. Hepatol.* **13**, 620 (2021).

Acknowledgements

This work has been supported (in part) by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project P2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation-Project Number 427806116).

Author contributions

J.R. conceived the original idea. C.D. performed the calculations with the help of H.K. J.D. wrote the manuscript with input from all authors. J.R. and F.K. supervised the work, revised the manuscript, and proposed analyses in discussions with J.D. and C.D. All authors read and improved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47057-0.

Correspondence and requests for materials should be addressed to J.C.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article 4

Bayesian non-linear subspace shrinkage using horseshoe priors

Julia Christin Duda *duda@statistik.tu-dortmund.de*
Department of Statistics, TU Dortmund University, Dortmund, Germany

Matthew Wheeler *matt.wheeler@nih.gov*
Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, Durham, North Carolina, U.S.A.

2024/06/05

Abstract

When modeling biological responses using Bayesian non-parametric regression, prior information may be available on the shape of the response in the form of non-linear function spaces that define the general shape of the response. To incorporate such information into the analysis, we develop a non-linear functional shrinkage (NLFS) approach that uniformly shrinks the non-parametric fitted function into a non-linear function space while allowing for fits outside of this space when the data suggest alternative shapes. This approach extends existing functional shrinkage approaches into linear subspaces to shrinkage into non-linear function spaces using a Taylor series expansion and corresponding updating of non-linear parameters. We demonstrate this general approach on the Hill model, a popular, biologically motivated model, and show that shrinkage into combined function spaces, i.e., where one has two or more non-linear functions a priori, is straightforward. We demonstrate this approach through synthetic and real data. Computational details on the underlying MCMC sampling are provided with data and analysis available in an online supplement.

1 Introduction

When modeling complex biological systems, mechanistic knowledge about the system under investigation is often available; however, including this information in a statistical model may be impossible due to the system's complexity in relation to experimental and computational resources [Mesarovic et al., 2004]. Often, simplified models are used in lieu of the true mechanistic model [Šimon, 2005]. When using these simplified models, one expects them to describe the observed data correctly or be mildly misspecified, and in the case of misspecification, the model may still be helpful in describing the response.

When modeling biological systems, an example of this situation is the use of the Hill model. This model, which represents sigmoidal-shaped responses, is a simplification of the complex biochemical process based upon chemical kinetics [Hill, 1910] and is used to model a wide variety of biochemical processes [Goutelle et al., 2008]. Despite its widespread use, it may not always represent the observed response. Non-monotone deviations of the Hill's functional form may be evident in the data. Additionally, other competing models may also be available, and the modeler might like to include this information to inform the fitting process, too. We develop a framework that allows one to define a subspace over one or more function spaces of interest for Bayesian non-parametric regression.

From the Bayesian perspective, there is a rich literature on approaches incorporating prior knowledge in non-parametric regression. Naively, one may center the non-parametric model on the specified parametric function. When the parametric data-generating mechanism's mean is the known parametric model, ensuring that estimates do not contain artifactual deviations from that model is difficult, implying that shrinkage to the prior model will not be uniform. Further, using this method, there is no way to create a space based on multiple parametric functions. More sophisticated approaches use shape constraints induced through the prior distribution, which include monotonicity or limits to the number of extrema [Brezger and Steiner, 2008, Shively et al., 2009, Meyer, 2008, Shively et al., 2011, Meyer et al., 2011, Gunn and Dunson, 2005, Köllmann et al., 2014, Wheeler et al., 2017]. Though these approaches are often effective, they do not directly incorporate parametric modeling information on the shape of the model; they force the response to be in the constrained space by putting a prior mass of zero on all responses outside of that space.

Alternatively, one may merge mechanistic prior knowledge into a model is through ordinary differential equations (ODEs) within a Bayesian framework. Parametric Bayesian models include pharmacokinetic/pharmacodynamic modeling, discussed by Lunn et al. [2002], and Huang et al. [2006] present an HIV-modeling example using Bayesian hierarchical models with non-linear differential equations. More flexible non-parametric approaches use differential equations to inform stochastic processes with induced constraints [Golightly and Wilkinson, 2011, Titsias et al., 2012]. While Alvarez et al. [2013] and Wheeler et al. [2014] proposed a Gaussian Process (GP) approach that incorporates mechanistic knowledge defined by differential equations. More recently, Chen et al. [2022] incorporate mechanistic knowledge defined by linear or non-linear partial differential equations (PDE) into a GP framework by selecting PDE points, i.e. pseudo covariate points through which the assumed PDE information is incorporated. Like the shape-constrained approaches, these methods form a Basis expansion consistent with a subspace defined using mechanistic knowledge. Thus, these priors imply that an estimated function is within the given subspace, and they do not allow for deviations outside of this space.

We define a prior distribution over a non-linear subspace - such as the Hill model and power model - that does not require a fitted function to be within that subspace. When the non-linear subspace is correctly specified, shrinkage into it occurs; but, when the true model is outside of the subspace, the approach is unconstrained. We build upon the work of Shin et al. [2020] who introduced the functional horseshoe (fHS) prior for linear spaces. The fHS prior shrinks the non-parametric fit towards a pre-specified, linear subspace. This approach is different from well-known shrinkage approaches such as Ridge, Lasso or Horseshoe [Hoerl and Kennard, 1970, Tibshirani, 1996, Carvalho et al., 2010], which shrink model coefficients in a non-parametric regression towards the origin. The prior of Shin et al. [2020] has the appealing property that the posterior shrinks into the pre-specified subspace f it is consistent with the observed data or, alternatively, is left unconstrained otherwise. The shrinkage occurs at the minimax optimal rate.

In our extension, we use a Taylor expansion to locally linearize the response function, where the derivatives depend on parameters of the non-linear model. The extension allows functional shrinkage into a non-linear function space or adapts the function to be outside of the non-linear space. The relevant non-linear function space is specified *a priori* using one or more parametric models.

We present our shrinking approach in Section 2. Section 3 then illustrates the approach both for the case of shrinkage into a single function space - shown for the Hill model - and into a combined function space - shown for the Hill and the power models. We compare our method against other parametric and non-parametric approaches in a simulation study in Section 4. We apply our method to a real-world data example of total testosterone levels measured in 9943 males aged between 3 and 85 years in section 5. The computational back-bone of the approach is MCMC sampling combining Gibbs-, Metropolis-Hastings- and Slice-sampling [Brooks et al., 2011, Neal, 2003], detailed in the supplementary material.

2 Model

2.1 Spline Model

Consider the non-parametric regression problem

$$y_i = g(x_i) + \varepsilon_i, \tag{1}$$

with unknown mean function $g : \mathbb{R} \rightarrow \mathbb{R}$. We observe $y = (y_1, \dots, y_n)'$ corresponding to covariates $x = (x_1, \dots, x_n)'$ and wish to estimate g. Assuming $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, it is common to approximate g using a B-spline basis expansion [Carl, 2001], i.e.,

$$f(x_i) = \sum_{m=1}^{k} \phi_m^j(x_i)\beta_m, \tag{2}$$

or $f(x) = \Phi\beta$. Here, the B-spline basis $\phi_k^j(x)$ are of order j defined on k^* internal knots, where $k = k^* + j$, $\beta = (\beta_1, \dots, \beta_k)^\top$ denotes the vector of basis coefficients. We consider cubic splines and omit the superscript $j = 3$. With a dense knot set, the spline approximation f can be made to be arbitrarily close to any continuous g, allowing one to estimate a large space of functions to arbitrary precision.

2.2 Bayesian Priors for a Non-linear Subspace

For many prior specifications, the expansion in (2) may not place high prior probability on biologically relevant responses. To define a biologically relevant model, we construct a prior distribution that places significant prior mass on the function space defined by the non-linear model, e.g., the space of Hill models, but does not put zero mass outside the function space.

To do this, assume knowledge about the shape of g through a twice differentiable function $h_\theta : \mathbb{R} \rightarrow \mathbb{R}$. The function h_θ depends on parameter vector θ, and defines the function space $\Omega_0^\Theta = \{h_\theta | \theta \in \Theta\}$ for all realizations $\Theta \subseteq \mathbb{R}^s$. If the true mean function g happens to be outside Ω_0^Θ, shrinkage towards Ω_0^Θ is undesirable. Given a dense knot set, the spline f can approximate h_θ for any $\epsilon-$ball. Consequently, the space of functions represented by the spline contains Ω_0^Θ. We define a prior for (2) that places prior mass on Ω_0^Θ, but does not limit responses to be only in Ω_0^Θ.

To define this prior, we consider Shin et al. [2020], who defined a projection prior that shrinks into the linear column space defined by the matrix $\Phi_0 \in \mathbb{R}^{n \times d}$ through

$$p(\beta|\sigma^2, \tau^2) \propto (\tau^2)^{-(k-d_0)/2} \exp\left(-\frac{1}{2\sigma^2\tau^2}\beta^\top\Phi^\top(I - P_{\Phi_0})\Phi\beta\right), \tag{3}$$

where $d_0 = \text{rank}(\Phi_0)$, Φ_0 is constructed as a linear space of known covariates, and P_{Φ_0} is the orthogonal projection matrix into the column space of Φ_0. The hyperparameter $,\tau,$ is given a generalized horseshoe (HS) prior with hyperparameters a and b (cf. Shin et al. [2020]). When $a = b = 0.5$ the prior is a half-Cauchy distribution, and one arrives at the HS prior [Carvalho et al., 2010].

In (3), one constructs P_{Φ_0} using the linear column space of Φ_0. Given our space is non-linear, there is no direct analogue to P_{Φ_0}. As an approximation, we use a Taylor series approximation of

h_{θ_0}, $\theta_0 \in \Theta$. That is, we linearly approximate Ω_0^Θ at any θ_0 using a first-order Taylor expansion

$$h_\theta(x) \approx h_{\theta_0}(x) + \dot{\mathsf{H}}_{\theta_0}(x)(\theta - \theta_0) \tag{4}$$

where $\dot{\mathsf{H}}_{\theta_0}(x) = \frac{\partial h_\theta(x)}{\partial \theta}\Big|_{\theta=\theta_0} \in \mathbb{R}^{n \times s}$ is the Jacobian containing the partial derivatives of h_θ evaluated at θ_0. The column space of $\dot{\mathsf{H}}_\theta(x)$ approximates $h_\theta(x)$ [Seber and Wild, 2003][p. 130] and we use $\dot{\mathsf{H}}_\theta(x)$ to construct $P_\theta = P_{\dot{\mathsf{H}}_\theta}$. Thus, for any θ_0, we project $f(x)$ onto the space locally approximating h_{θ_0}. When there are multiple function spaces to consider, the same approach applies; in this case, operator $P_{\dot{\mathsf{H}}_\theta}$ defines the projection into a combined linear space, where $\dot{\mathsf{H}}_\theta$ represents the Jacobian across all assumed functions.

We place the prior

$$p(\beta|\sigma^2, \tau^2, \theta) \propto (\tau^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2\tau^2}\beta^\top\Phi^\top(I - P_\theta)\Phi\beta\right) \tag{5}$$

over β to shrink realizations of (2) into Ω_0^Θ. In (5), θ is given an appropriate prior to complete the specification. This approach penalizes deviations of $\Phi\beta$ based upon the projection operator $(I - P_\theta)$. As we shrink back to a planar approximation given a specific θ_0, we require appropriately specified priors for the non-linear parameters in Θ. As β is defined conditional on θ through the linear projection operator P_θ, only priors for the non-linear parameters can be learned.

In this formulation, $(\tau^2)^{-(k-d_0)/2}$ in (3) becomes $(\tau^2)^{-k/2}$ because we separately model the intercept (cf. Section 3.1). This change yields proper priors as due to the non-linearity, no linear basis of Φ is in $(I - P_\theta)$ and $\Phi^\top(I - P_\theta)\Phi$ has full rank.

3 Non-linear functional shrinkage for single or combined function spaces

3.1 Single function spaces

As an example of non-linear functional shrinkage using a single function, we consider the Hill model. This function is given by

$$h(x, \theta) = \theta_1 + \theta_2 \frac{x^{\theta_4}}{{\theta_3}^{\theta_4} + x^{\theta_4}}, \tag{6}$$

where θ_1 is the background response at $x = 0$, θ_2 is the maximal change in the response, θ_3 is the dose where half of this change is reached and θ_4 defines the steepness of the curve. The Jacobian, $\dot{\mathsf{H}}_\theta$, is

$$\frac{\partial h(x,\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} = \begin{pmatrix} 1 & \frac{x^{\theta_4}}{x^{\theta_4} + {\theta_3}^{\theta_4}} & \theta_2 \frac{-\theta_4}{\theta_3} s(x, \theta_3, \theta_4) & \theta_2 \log(\theta_3/x) s(x, \theta_3, \theta_4) \end{pmatrix}\bigg|_{\theta=\theta_0}, \tag{7}$$

with $s(x, \theta_3, \theta_4) = ((x{\theta_3}^{-1})^{\theta_4} + 1)^{-1}((\theta_3 x^{-1})^{\theta_4} + 1)^{-1}$. The derivative matrix does not depend upon the linear parameters θ_1, but it still depends on θ_2. However, P_θ does not depend on θ_1 and θ_2 (cf. Lemma 1 in the appendix), which gives a direct example of why we do not place a prior over these linear quantities. Of the parameters in (7), parameter θ_3 is of particular interest because it represents the value of x that produces a response that is the average of the lower and upper asymptote. Values of the covariate below θ_3 correspond to values of the response less than 50% of the maximal response. Further, θ_4 corresponds to the steepness of the response and speed of a chemical reaction in a biological substrate. As both quantities have direct interpretation, informative priors can be developed for these quantities accordingly, which in turn informs the subspace the model may shrink into.

To specify the hyperprior over (θ_3, θ_4), we assume $x \in [0, 1]$, and let $E[\theta_3] = 0.5$, the midpoint, and for θ_4, we center it on 3, letting the parameter vary within a range that we have often seen in bioassays. In our model, θ_1 enters as the intercept, and θ_2, the maximal response change, implicitly enters the model through the β coefficients. Using the Hill model as a prior to define (5), we complete the prior specification as

$$(y|\beta, \sigma^2, \theta_1) \sim N(\theta_1 + \Phi\beta, \sigma^2 I_n) \tag{8}$$

$$\theta_1 \sim N(0, 20), \quad \theta_3 \sim N_+(0.5, 0.05) \quad \theta_4 \sim LN(0.95, 0.29) \tag{9}$$

$$\sigma^2 \sim IG(0.001, 0.001), \tag{10}$$

where $N_+(a, b)$ is a truncated normal distribution with mean a and variance b (before truncation), $LN(a, b)$ is a log-normal distribution with log-mean a and log-variance b, and $IG(a, b)$ is an inverse-gamma distribution with shape a and scale b. Note that $\theta_4 \sim LN(0.95, 0.29)$ results in $E[\theta_4] = 3$ and $V[\theta_4] = 3$.

3.2 Combined function spaces

If one desires multiple functions to define in the function space because of uncertainty in the function space, one can add multiple functions. Here, assume there are $r \in \{1, \ldots, R\} = \mathcal{R}$ function spaces $\Omega_0^{(r)} = \{h_\theta^{(r)} | \theta \in \Theta^{(r)}\}$ of interest; we omit the index r on each θ for simplicity. For each $\Omega_0^{(r)}$, calculate the Jacobian, $\dot{\mathsf{H}}_\theta^{(r)}$, i.e.,

$$\dot{\mathsf{H}}_\theta^{(\mathcal{R})} = (\dot{\mathsf{H}}_\theta^{(1)} \ldots \dot{\mathsf{H}}_\theta^{(R)}),$$

and use this to construct P_θ. The Jacobian, $\dot{\mathsf{H}}_\theta^{(\mathcal{R})}$, must be full rank without linear bases other than an intercept column for Equ. 5 to hold.

To illustrate the combined subspace shrinkage approach, we use the Hill and power models. The latter function defined as as $h_\theta(x) = \theta_1 + \theta_2 x^{\theta_3}$, which has only one non-linear parameter,

θ_3, that requires a prior specification. We use $\theta_3 \sim N(0.5, 0.25)$, to center on a concave shape. The partial derivatives of the power model are

$$\frac{\partial h(x,\theta)}{\partial \theta} = \begin{pmatrix} 1 & x^{\theta_3} & \log(x)x^{\theta_3} \end{pmatrix} = \dot{\mathsf{H}}_{\theta}^{(1)}. \tag{11}$$

Prior to combining $\dot{\mathsf{H}}_{\theta}^{(1)}$ of the power model and $\dot{\mathsf{H}}_{\theta}^{(2)}$ of the Hill model (Eq. 7), we remove the intercept from $\dot{\mathsf{H}}_{\theta}^{(1)}$ to obtain a full rank. Shrinkage into the combined subspaces is equivalent to shrinkage into a single subspace.

4 Simulation Study

4.1 Setup

We perform a simulation study and evaluate the performance of the proposed approach against other fitting strategies. Full details of the simulation design are summarized by the ADEMP principle described in Morris et al. [2019] (Table S2). We generate data using three parametric cases: the Hill model, the power model, and a misspecified model (the Hill model with downturn). We look at exposure-response data as, for such data, chemical kinetics of exposure are often approximated by the Hill model, but the results generalize to other domains.

For each data set, we draw $x \in [0, 1]$ uniformly for $n \in \{50, 100, 200, 500\}$ observations, where 50 is a realistic assay size and larger n are chosen to study the large sample behavior. Mean zero normal noise with variance $\sigma^2 = 0.005$ and a larger noise $\sigma^2 = 0.05$ is added. These variances represent a 2-SD spread that is approximately 14% and 45% of the maximal response. In total, 24 data generating scenarios are used, with $n_{\text{rep}} = 1000$ simulations per scenario. For each dataset, we apply the following methods:

4.1.1 Modeling Approaches

Non-linear functional shrinkage (NLFS)

Non-linear functional shrinkage is performed with shrinkage into the Hill space (NLFS(Hill)), power space (NLFS(power)), or a combination (cf. Section 3.2) of both function spaces (NLFS(Hill+power)). Two variations for the shrinkage parameter τ^2 are considered. One uses a half Cauchy prior ($a = b = 0.5$) and is implemented according to Makalic and Schmidt [2015]; the other, implemented ourselves using slice sampling [Neal, 2003], uses a $\omega \sim \text{Beta}(a, b)$ prior where $a = 0.5$ and $b = \exp(-k \log(n)/2)$ as proposed by Shin et al. [2020] and k is the number of knots.

Parametric Model (Param.)

We investigate the performance fitting of the Oracle model using Bayesian parametric regression for the Hill model (Param.(Hill)) or the power model (Param.(power)) (priors in Table S2). Fitting these models allows us to compare the performance of the Oracle NLFS to the Oracle parametric model.

B-splines

Bayesian B-splines with a scaling parameter $\lambda^2 \sim \text{IG}(0.001, 0.001)$ where the spline coefficients are given by the prior $\beta \sim N(0, \sigma^2\lambda^2\text{diag}(k))$. This model represents a basis approach without smoothing and is used to compare the performance of the NLFS approach when the shrinkage subspace is misspecified.

P-splines

Penalized Bayesian smoothing splines where $\beta \sim N(0, \sigma^2\tau^2K^{-1})$ where $K = R^\top R$ and R is a second order penalty matrix and $\tau^2 \sim \text{IG}(1, 0.005)$, similar to the hyperparameter choices in Lang and Brezger [2004]. This approach builds upon the B-spline approach, adding a smoothing component, and typically performs better in practice than B-splines

Parametric Model + horseshoe B-spline

We also consider a model that includes the true parametric model plus an additional B-spline to account for model misspecification, i.e., $y = h_\theta(x) + \Phi\beta + \varepsilon$. When $h_\theta(x)$ specifies the correct model, one has $\beta = 0$; otherwise, $\beta \neq 0$. To shrink the β coefficients to zero, we use a horseshoe prior, i.e., $\beta \sim N(0, \sigma^2\tau^2\text{diag}({\lambda_1}^2, \ldots, {\lambda_k}^2))$ where $\tau \sim C^+(0, 1)$ and $\lambda_j \overset{\text{iid}}{\sim} C^+(0, 1)$, cf. Makalic and Schmidt [2015]. $C^+(0, 1)$ denotes a standard Half-Cauchy prior. As in the parametric model case, h_θ is either the parametric Hill (Param.(Hill)+B-spline) or power model (Param.(power)+B-spline). This approach represents a direct competitor to the NLFS approach.

4.1.2 Further Considerations

For all simulations, we use $k = 15$ inner knots for the B-spline basis matrix. When MCMC sampling, we took 10,000 draws from the posterior, discarding the first 2000 samples as burn-in. Initial experiments indicated that this number of samples was adequate to estimate the posterior distribution. For the spline-based approaches (NLFS, B-spline, P-spline), we place a vague prior on the intercept term, θ_1, defined in (8). Traceplots, of an NLFS fit with correctly and incorrectly specified subspaces, are given in the supplement (supplemental Figures 4 and 5) and show convergence.

4.2 Results

Figure 1 gives representative results of the simulation, where all results are provided in the supplement. Unsurprisingly, when the Hill model is the truth (Figure 1a), we observe the largest RMSE of 0.151 when fitting the misspecified parametric model (Param.(power)). Unlike the misspecified parametric fits, when the function space is misspecified in the NLFS approach (NLFS(power)), the RMSE is approximately one-third (0.046) that from fitting the misspecified parametric model, indicating the NLFS approach adjusts to the data. In this scenario, the NLFS(power) performance with mean RMSE of 0.046 was similar to that of the B-spline approach with mean RMSE of 0.042.

When the correct function space is assumed for the NLFS prior, the RMSE drops to 0.019 (Figure 1a), as low as that of the 0racle parametric fit. This demonstrates the adaptive shrinkage behavior of the NLFS approach in the case of correct subspace specification. Here, the NLFS approach effectively shrank towards the correctly assumed space for sample sizes as low as $n = 50$, and performed similarly to the oracle parametric model fit. The P-spline approach receives no prior model or subspace specification but yields smooth splines. Consequently, its performance was in between the approach with misspecification and correct specification.

When the correct function space is assumed, the NLFS approach tended to outperform the parametric + horseshoe B-spline (PHBspline) approach. The PHBspline approach does not enforce an equally smooth, global shrinkage of all β towards zero, especially when there are leverage points far from the observed mean.

When all assumed models or spaces are misspecified (Figure 1b), the NLFS approach was outperformed by the PHBspline approach for the same model misspecification, i.e., NLFS(power) was outperformed by param.(power) + horseshoe B-spline and NLFS(Hill) was outperformed by param.(Hill) + horseshoe B-spline. However, the NLFS approach in general has an advantage

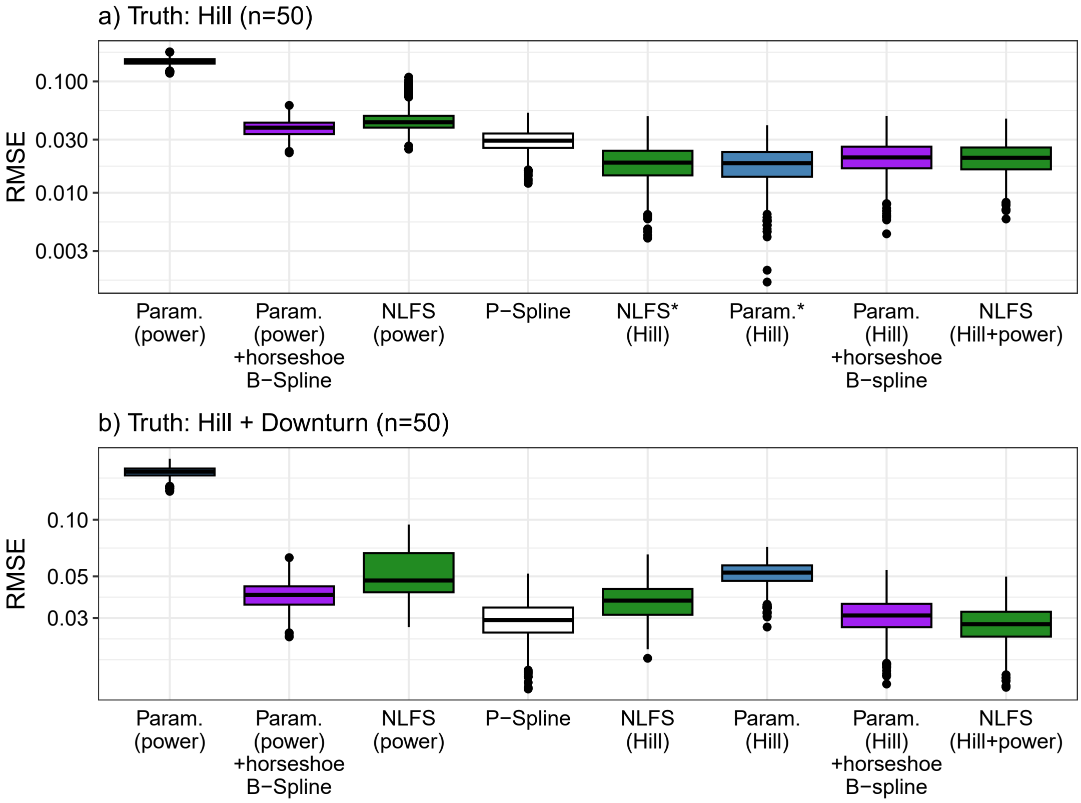


Figure 1: Representative root mean square error (RMSE) results of the simulation for the scenarios where the truth is Hill (a) or Hill + Downturn (b) and a medium noise level. Pane b represents the situation where a deviation of an unknown shape is the truth. All simulation results can be found in Tables S3 and S4. *Correct model used in the model fit.

in misspecification scenarios due to its inherent flexibility to shrink toward combined function spaces. The NLFS(Hill+power) outperformed all other approaches in this scenario with a mean RMSE of 0.028. Only the P-spline approach came close, showing a slightly weaker performance with a mean RMSE of 0.030.

The NLFS prior appropriately shrinks into the correct space, giving equivalent fits to the parametric Hill model (Figure 2a). Correspondingly, the P-Spline smoothing approach estimate shows various artifactual bumps not evident in the NLFS approach. Even if the true model is the Hill model, the naive PHBspline approach produces an artifactual bump in the asymptote region that does not occur with the NLFS fit (Figure 2d). The NLFS approach and generic B-spline are equivalent when the model is misspecified (i.e., fits a model outside of the assumed space) (Figure 2b). The NLFS(Hill+power) approach fits a model outside of the Hill space (Figure 2c) and illustrates how shrinkage into combined subspaces can reduce misspecification errors involving minor deviations.

5 Real Data Example

We applied the proposed method on a testosterone data set collected by Kelsey et al. [2014]. They modeled total testosterone (TT) concentration in male participants dependent on age, to identify normal TT ranges at any age. TT levels are the result of highly complex physiological processes, mechanistic models are not available. TT is expected to increase during puberty, reach a maximum and possibly slowly decline with age, which is a sigmoidal assumption. Consequently, we *a priori* assume the Hill based shape through the NLFS prior.

Kelsey et al. [2014] collected data from 13 studies on TT by age, yielding $>$10.000 data points; they then fit 330 polynomial models and selected a single best parametric model based on the best R^2 with 5-fold cross-validation. Due to the large number of data points, spline smoothing approaches tend to produce local artifacts that are biologically unreasonable. The NLFS approach offers an alternative to extensive model comparisons while simultaneously incorporating knowledge on the curves shape.

For the analysis, we set

$$\theta_3 \sim N(15, 4), \tag{12}$$

$$\theta_4 \sim LN(2.28, 0.05). \tag{13}$$

Since testosterone levels increase during puberty, *a priori* we assume TT levels to reach half maximal levels (e.g., θ_3) at approximately 15 years with a standard deviation of 2 years. For the steepness parameter, θ_4, we set the log-mean to 2.28 and the log-variance to 0.05 implying that

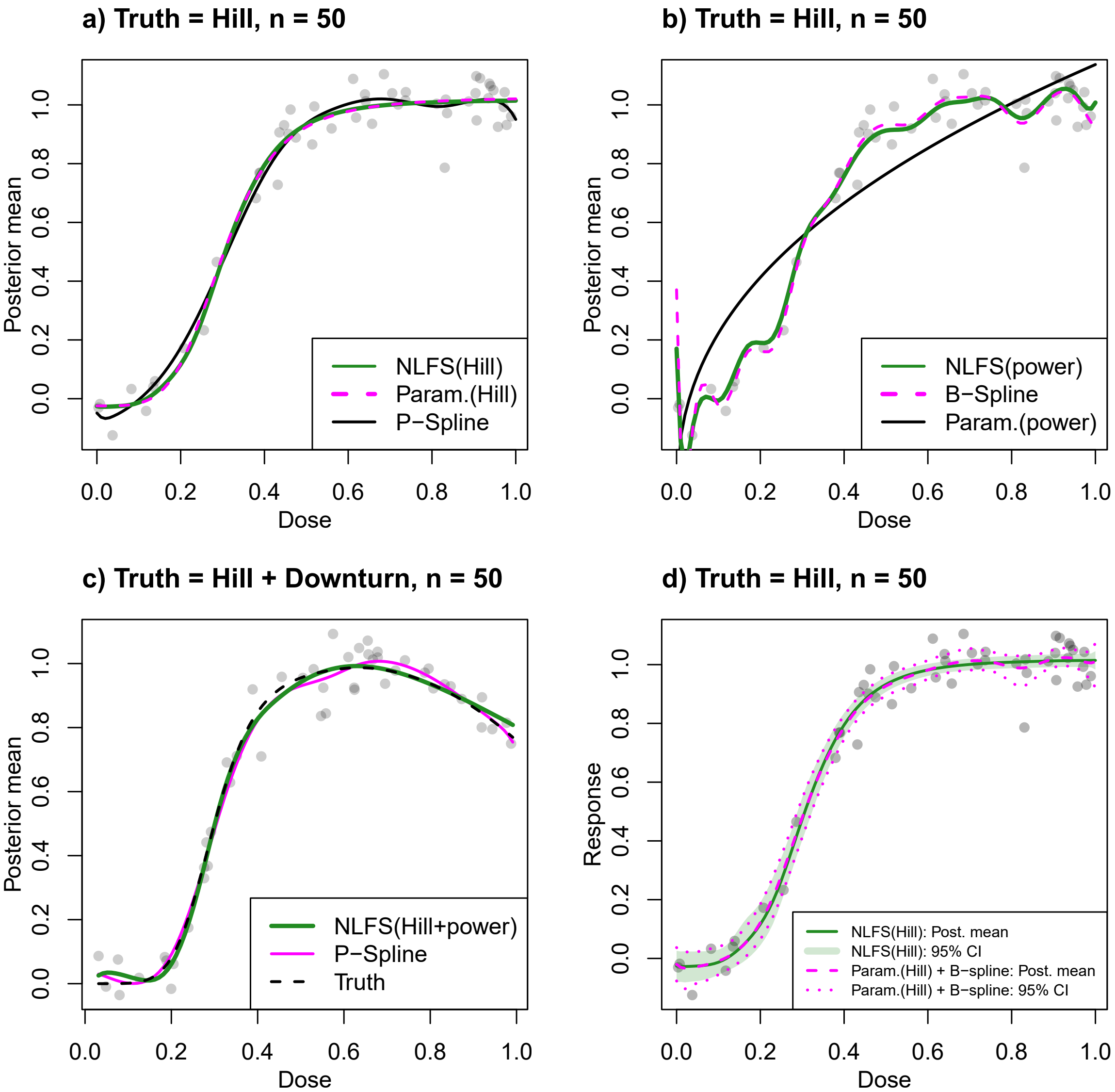


Figure 2: Posterior mean responses (a-d) and credible intervals (d) of representative simulation runs with median noise level ($\sigma^2 = 0.005$): (a) Oracle scenario. (b) misspecification scenario. (c) deviations of unknown shape. Combined subspace shrinkage reduces misspecification errors. (d) Comparison of credible intervals between NLFS and PHBspline in the Oracle scenario.

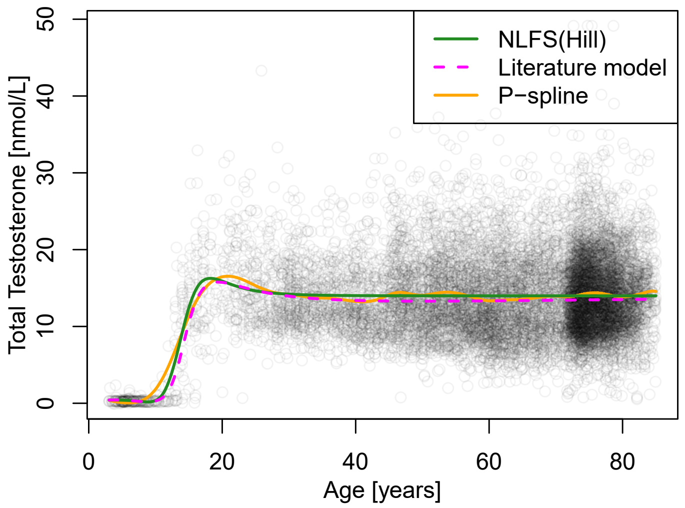


Figure 3: Comparison of the NLFS approach to a P-spline fit and the best parametric model selected by Kelsey et al. [2014]. For the P-spline and NLFS(Hill), the posterior mean is displayed. The parametric model by Kelsey is $\log_{10}(TT+1) = (a + cx + ex^2 + gx^3)/(1 + bx + dx^2 + fx^2)$ with $a = 0.04655$, $b = -0.05311$, $c = 0.05123$, $-d = 0.00793$, $e = -0.01222$, $f = 0.00058$ and $g = 0.00069$. For display, the model is backtransformed using $\exp(\log_{10}(TT+1) + \hat{\sigma}^2/2)$ where $\hat{\sigma}^2$ is the estimated residual variance.

the actual mean and variance to be 10 and 5, respectively. We choose these values as we expect TT to increase rapidly upon the onset of puberty.

We fit our model to the slightly reduced data set of men of age 85 or younger (98.5% of original data) because of extreme variability in the approximately 150 observations above 85. To model the noise, we let $\sigma^2 \sim C^+(0,1)$, to account for the large variance in the data.

The NLFS and parametric fit by Kelsey et al. [2014] are roughly sigmoidal (Figure 3). They both show a peak around age 19, followed by a slight descent that eventually plateaus. The NLFS approach notably predicts a larger mean testosterone level than the parametric model. The P-spline fit has a less pronounced peak TT around age 20 and has artifactual bumps that oscillate around the NLFS estimate. For each method, the observed RMSE values were 5.123 (NLFS), 5.154 (parametric), and 5.111 (P-spline) and therefore similar, but the three resulting model fits are visibly distinguishable. Arguably, the NLFS approach more appropriately models the mean than the P-spline due to the latter's bumpiness. Further, NLFS estimates a higher mean TT level than the parametric model, which suggests there may be some underestimation of the mean TT when using a parametric approach.

6 Discussion

The NLFS prior enables adaptive shrinkage into a pre-specified non-linear function space but does not constrain the resulting function to be in that space if the data are not compatible with that *a priori* function space. This approach can be applied to function spaces defined by any twice differentiable function. Because such a setting is common, the NLFS approach balances adhering to prior assumptions and accounting for model misspecification.

The NLFS approach can shrink into a combined function space, thereby providing robustness against misspecification. This benefit is supported by simulation results, where NLFS combined function space prior outperformed all other methods under model misspecification. Defining such a prior is straightforward for the NLFS approach. Attempting to account for parametric misspecification by including a spline that shrinks to zero if the model is correctly specified can give artifactual features in the Oracle scenario. As a comparison, the NLFS prior gave slightly better RMSE results under the Oracle scenario, and provided more realistic curve fits without the artifactual features. Though NLFS with a misspecified single subspace performed slightly worse, adding subspaces in NLFS did not lead to a relevant performance loss compared to only including the correct model but also robustified NLFS against misspecification.

When modeling the TT data in Kelsey et al. [2014], the NLFS yielded a plausible nonparametric estimate that did not produce artifactual features. In this regard, NLFS provided equally reasonable mean estimates as the parametric model, while not requiring a model selection procedure on over 300 models.

Simulation results empirically show that NLFS correctly decides to either shrink towards the specified subspace, or remain unconstrained. Though we have not provided a theoretical proof, our simulation results suggest that the optimal, theoretical shrinkage properties given by Shin et al. [2020] approximately hold in the non-linear case. Because it performs similarly to an unsmoothed spline estimate, adding a smoothness penalty similar to the one proposed by Wiemann and Kneib [2021] for linear subspace shrinkage may be a promising extension.

The construction of NLFS assumes the independent variable to be continuously distributed, with a unique covariate value for each observation. In some biological applications, data are generated in a planned experimental setting, with multiple units treated at few distinct exposure levels. For such experiments, the exposure is typically a dose or concentration. Dose-response modeling is often performed in terms of a simple parametric Hill model fit, which can lead to misspecification errors that could be prevented using NLFS. Tailoring NLFS to a such a data structure is necessary. This can be done using a grid that defines the shrinkage locations, such that the shrinkage is independent of the few experimentally selected doses. Precisely, Φ_θ would be evaluated at a grid instead of the observed exposure levels. This avoids a lack of shrinkage at basis functions that fall between exposure levels. This extension would yield more model parameters related to the construction of the grid. Another extension using a fully specified Gaussian processes is an alternative and would reduce hyperparameter choices on knot sequences and shrinkage grids. Another extension is to account for heteroscedasticity. For non-parametric Bayesian modeling, different methodologies can be applied, e.g. Dirichlet process priors. Other computational challenges in the NLFS approach relate to the derivatives. For example, using the Hill model, derivatives w.r.t. the non-linear parameters can be almost linearly dependent. Careful prior selection or expanding the shrinkage onto additional subspaces might soften this challenge.

Acknowledgements

This manuscript was funded in part by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation-Project Number 427806116) and intramural funds at the NIEHS.

Supporting Information

The code and data underlying this article are available on GitLab at
https://gitlab.tu-dortmund.de/functional_shrinkage/nonlinear_shrinkage.

A Tables

Table 1: Overview of method settings used in the simulation study.

	Algorithm	Assume	Shrinkage (Horseshoe prior)
1	NLFS	Hill	$a = b = 0.5$
2	NLFS	Hill	$a = 0.5,\ b = \exp(-k \log(n)/2)$
3	NLFS	Hill & power	$a = b = 0.5$
4	NLFS	Hill & power	$a = 0.5,\ b = \exp(-k \log(n)/2)$
5	NLFS	power	$a = b = 0.5$
6	NLFS	power	$a = 0.5,\ b = \exp(-k \log(n)/2)$
7	parametric	Hill	-
8	parametric	power	-
9	B-spline	-	-
10	P-spline	-	-
11	Parametric + B-spline	Hill	$a = b = 0.5$
12	Parametric + B-spline	power	$a = b = 0.5$

Table 2: Simulation setup summarized by the ADEMP principle.

Aim	Comparing proposed approach against existing approaches
Data generation	Dose-response models:
	- Hill: $h_\theta(x) = \frac{x^{\theta_4}}{\theta_3^{\theta_4}+x^{\theta_4}}$ ($\theta_3 = 0.3, \theta_4 = 6$)
	- Power: $h_\theta(x) = \theta_1 x^{\theta_2}$, ($\theta_2 = 0.5$)
	- Hill + Downturn: $h_\theta(x) = h_\theta^{\text{Hill}}(x) + \mathbb{1}_{[0.6,\infty)}(x)(-1.5(x-0.6)^2)$ ($\theta_3 = 0.3, \theta_4 = 6$)
	Doses: Unif$\sim [0,1]$
	Sample sizes: $n \in \{50, 100, 200, 500\}$
	Added noise: $\varepsilon \sim N(0,\sigma^2)$, $\sigma^2 \in \{0.005, 0.05\}$
Estimand	Mean of posterior dose-response function estimate $f(x)$
Methods	**Non-linear functional shrinkage** (NLFS) ($\theta_1 \sim N(0,1)$, $\sigma^2 \sim$ IG$(0.001, 0.001)$)
	- assuming Hill (NLFS (Hill))
	Priors: $\theta_3 \sim N_+(0.5, 0.05)$, $\theta_4 \sim LN$ s.t. $\mathbb{E}(\theta_4) = 3$, $\mathbb{V}ar(\theta_4) = 3$
	- assuming power (NLFS (power))
	Priors: $\theta_3 \sim N(0.5, 0.25)$
	- assuming Hill and power (NLFS (Hill+power))
	Priors: As in NLFS(Hill) and NLFS(power)
	Parametric Bayesian fit (Param.) ($\theta_1 \sim N(0,1)$, $\log(\sigma^2) \sim N(-1.75, 1)$)
	- assuming Hill (Param.(Hill))
	Priors: $\theta_3 \sim N_+(0.5, 0.05)$, $\theta_4 \sim LN$ s.t. $\mathbb{E}(\theta_4) = 3$, $\mathbb{V}ar(\theta_4) = 3$
	- assuming power (Param.(power))
	Priors: $\theta \sim N(0.5, 0.25)$
	B-spline
	Priors: $\theta_1 \sim N(0,1)$, $\sigma^2 \sim$ IG$(0.001, 0.001)$, $\lambda^2 \sim$ IG$(0.001, 0.001)$
	P-spline
	Priors: $\theta_1 \sim N(0,1)$, $\sigma^2 \sim$ IG$(0.001, 0.001)$, $\tau^2 \sim$ IG$(1, 0.005)$
	Parametric + horseshoe B-spline
	$y = h_\theta(x) + \Phi(\beta) + \varepsilon$
	Priors:
	$\beta \sim N(0, \sigma^2\tau^2$diag$(\lambda_1^2, \ldots, \lambda_k^2))$
	$\tau \sim C^+(0,1)$, $\lambda_j \overset{\text{iid}}{\sim} C^+(0,1)$
	$\theta_1 \sim N(0,1)$, $\theta_2 \sim N(1.5, 2)$ (Scaling)
	- assuming Hill (Param.(Hill) + B-spline))
	Prior: $\theta_3 \sim N_+(0.5, 0.05)$, $\theta_4 \sim LN$ s.t. $\mathbb{E}(\theta_4) = 3$, $\mathbb{V}ar(\theta_4) = 3$
	- assuming power (Param.(power) + B-spline)
	Prior: $\theta \sim N(0.5, 0.25)$
Performance	RMSE between posterior mean $\mathbb{E}(f(x) \mid d_s^i)$ and true $g(x)$ evaluated at drawn doses $x \in [0,1]^n$

Table 3: Simulation results for $\sigma^2 = 0.005$ ($2\sigma = 14.1\%$ of maximal effect) summarized by mean RMSE and corresponding standard deviation in parenthesis. OS means own slice and refers to setting a and b for the shrinkage parameters as suggested in Shin et al. [2020] whereas HC means Half Cauchy and refers to the standard horseshoe prior with $a = b = 0.5$.

		n=50			n=100			n=200			n=500		
	Method	Hill	power	Hill+down	Hill	power	Hill+down	Hill	power	Hill+down	Hill	power	Hill+down
1	NLFS(Hill), OS	0.019	0.027	0.037	0.014	0.017	0.028	0.009	0.011	0.024	0.006	0.008	0.017
		(0.007)	(0.008)	(0.008)	(0.005)	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.004)
2	NLFS(power), OS	0.046	0.018	0.053	0.031	0.013	0.031	0.023	0.009	0.022	0.015	0.006	0.015
		(0.013)	(0.006)	(0.017)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)	(0.002)
3	NLFS(Hill+power), OS	0.021	0.022	0.028	0.015	0.015	0.023	0.01	0.011	0.018	0.007	0.007	0.011
		(0.007)	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
4	NLFS(Hill), HC	0.019	0.022	0.033	0.014	0.015	0.027	0.01	0.011	0.024	0.006	0.008	0.015
		(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.004)
5	NLFS(power), HC	0.046	0.017	0.054	0.03	0.012	0.03	0.021	0.009	0.022	0.013	0.006	0.013
		(0.015)	(0.006)	(0.019)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)	(0.002)
6	NLFS(Hill&power), HC	0.021	0.021	0.028	0.015	0.015	0.023	0.01	0.011	0.018	0.007	0.007	0.011
		(0.007)	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
7	param(Hill)	0.019	0.023	0.052	0.013	0.019	0.05	0.009	0.016	0.05	0.006	0.012	0.049
		(0.007)	(0.005)	(0.008)	(0.005)	(0.004)	(0.005)	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)	(0.002)
8	param(power)	0.151	0.015	0.18	0.152	0.011	0.18	0.154	0.008	0.181	0.154	0.005	0.182
		(0.011)	(0.006)	(0.011)	(0.008)	(0.005)	(0.008)	(0.005)	(0.003)	(0.006)	(0.003)	(0.002)	(0.004)
9	bspline	0.042	0.042	0.042	0.032	0.033	0.032	0.023	0.024	0.023	0.014	0.014	0.014
		(0.007)	(0.008)	(0.007)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
10	pspline	0.03	0.026	0.03	0.022	0.019	0.022	0.016	0.015	0.016	0.011	0.01	0.011
		(0.007)	(0.006)	(0.007)	(0.005)	(0.004)	(0.005)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
11	param(Hill)+bspline	0.021	0.022	0.031	0.015	0.017	0.024	0.011	0.014	0.018	0.007	0.01	0.012
		(0.007)	(0.006)	(0.006)	(0.005)	(0.004)	(0.005)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
12	param(power)+bspline	0.038	0.019	0.04	0.029	0.013	0.03	0.021	0.009	0.022	0.013	0.006	0.014
		(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)	(0.002)

Table 4: Simulation results for $\sigma^2 = 0.05$ ($2\sigma = 44.7\%$ of maximal effect) summarized by mean RMSE and corresponding standard deviation in parenthesis. OS means own slice and refers to setting a and b for the shrinkage parameters as suggested in Shin et al. [2020] whereas HC means Half Cauchy and refers to the standard horseshoe prior with $a = b = 0.5$.

		n=50			n=100			n=200			n=500		
	Method	Hill	power	Hill+down	Hill	power	Hill+down	Hill	power	Hill+down	Hill	power	Hill+down
1	NLFS(Hill), OS	0.059	0.081	0.076	0.043	0.058	0.063	0.03	0.039	0.051	0.019	0.022	0.036
		(0.021)	(0.021)	(0.017)	(0.015)	(0.015)	(0.012)	(0.01)	(0.011)	(0.009)	(0.007)	(0.007)	(0.007)
2	NLFS(power), OS	0.137	0.053	0.155	0.107	0.039	0.101	0.066	0.027	0.067	0.042	0.017	0.043
		(0.015)	(0.019)	(0.022)	(0.023)	(0.014)	(0.024)	(0.012)	(0.011)	(0.012)	(0.007)	(0.007)	(0.007)
3	NLFS(Hill+power), OS	0.064	0.066	0.072	0.046	0.046	0.055	0.032	0.031	0.041	0.02	0.02	0.03
		(0.021)	(0.021)	(0.019)	(0.015)	(0.015)	(0.014)	(0.01)	(0.011)	(0.01)	(0.007)	(0.007)	(0.006)
4	NLFS(Hill), HC	0.057	0.067	0.073	0.042	0.049	0.059	0.03	0.034	0.046	0.02	0.021	0.033
		(0.02)	(0.02)	(0.017)	(0.015)	(0.015)	(0.013)	(0.01)	(0.011)	(0.01)	(0.007)	(0.007)	(0.006)
5	NLFS(power), HC	0.133	0.048	0.141	0.098	0.036	0.097	0.066	0.026	0.072	0.042	0.017	0.043
		(0.016)	(0.02)	(0.026)	(0.021)	(0.015)	(0.019)	(0.013)	(0.011)	(0.014)	(0.007)	(0.007)	(0.008)
6	NLFS(Hill+power), HC	0.061	0.06	0.068	0.045	0.045	0.053	0.032	0.032	0.041	0.02	0.02	0.03
		(0.02)	(0.019)	(0.019)	(0.015)	(0.015)	(0.014)	(0.01)	(0.011)	(0.01)	(0.007)	(0.007)	(0.006)
7	param(Hill)	0.063	0.054	0.082	0.043	0.042	0.066	0.03	0.033	0.058	0.019	0.024	0.053
		(0.021)	(0.018)	(0.016)	(0.016)	(0.013)	(0.011)	(0.011)	(0.009)	(0.006)	(0.007)	(0.005)	(0.003)
8	param(power)	0.161	0.043	0.188	0.157	0.032	0.184	0.156	0.024	0.183	0.155	0.016	0.183
		(0.012)	(0.019)	(0.012)	(0.008)	(0.014)	(0.009)	(0.005)	(0.01)	(0.006)	(0.003)	(0.007)	(0.004)
9	bspline	0.112	0.107	0.111	0.088	0.084	0.086	0.068	0.067	0.067	0.047	0.048	0.046
		(0.02)	(0.022)	(0.02)	(0.014)	(0.014)	(0.014)	(0.01)	(0.01)	(0.01)	(0.006)	(0.006)	(0.006)
10	pspline	0.081	0.063	0.08	0.061	0.048	0.06	0.045	0.037	0.045	0.03	0.026	0.03
		(0.019)	(0.017)	(0.019)	(0.014)	(0.013)	(0.013)	(0.01)	(0.009)	(0.01)	(0.006)	(0.006)	(0.006)
11	param(Hill)+bspline, HC	0.072	0.06	0.081	0.05	0.045	0.061	0.034	0.034	0.047	0.022	0.023	0.033
		(0.021)	(0.02)	(0.019)	(0.016)	(0.014)	(0.014)	(0.011)	(0.01)	(0.009)	(0.007)	(0.006)	(0.006)
12	param(power)+bspline, HC	0.097	0.053	0.103	0.076	0.039	0.081	0.058	0.028	0.061	0.04	0.019	0.042
		(0.019)	(0.02)	(0.019)	(0.014)	(0.015)	(0.014)	(0.01)	(0.01)	(0.01)	(0.006)	(0.007)	(0.006)

B Figures

C Proofs

Lemma 1: P_θ does not depend on linear parameters.

Let $h(x,\theta) = \theta_1 + \theta_2 q(x,\theta_3)$ be a twice differentiable function with q non-linear in θ_3. W.L.o.G. assume that $\theta_3 \in \mathbb{R}$. Then

$$\dot{\mathsf{H}}_\theta = \frac{\partial h(x,\theta)}{\partial \theta} = (1_n \quad \underbrace{q(x,\theta_3)}_{:=c_1} \quad \theta_2 \underbrace{\frac{\partial q(x,\theta_3)}{\partial \theta_3}}_{:=c_2}) = \underbrace{(1_n \quad c_1 \quad c_2)}_{:=H_1 \in \mathbb{R}^{n\times 3}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \theta_2 \end{pmatrix}}_{H_2 \in \mathbb{R}^{3\times 3}}$$

and

$$\begin{aligned} P_\theta &= \dot{\mathsf{H}}_\theta (\dot{\mathsf{H}}_\theta^\top \dot{\mathsf{H}}_\theta)^{-1} \dot{\mathsf{H}}_\theta^\top \\ &= H_1 H_2 (H_2 H_1^\top H_1 H_2)^{-1} H_2 H_1^\top \\ &= H_1 H_2 H_2^{-1} (H_1^\top H_1)^{-1} H_2^{-1} H_2 H_1^\top \\ &= H_1 (H_1^\top H_1)^{-1} H_1^\top \end{aligned}$$

and H_1 does not depend on θ_2.

D Computation

The code to reproduce results is available at
https://gitlab.tu-dortmund.de/functional_shrinkage/nonlinear_shrinkage.

The non-linear functional shrinkage (NLFS) approach for the Hill model is implemented using a combination of Gibbs-, Metropolis-Hastings- and Slice sampling [Brooks et al., 2011, Neal, 2003]. We separately model the function intercept θ_1.

Given the likelihood and priors

$$\begin{aligned} &Y \sim N(\theta_1 1_n + \Phi\beta, \sigma^2 I_n) \\ &\sigma^2 \sim \mathsf{IG}(a_\sigma, b_\sigma),\ \theta_1 \sim N(\mu_{\theta_1}, \sigma^2_{\theta_1}) \\ &\beta \sim N(0, \sigma^2\tau^2(\Phi^\top(I - P_\theta)\Phi)^{-1}) \\ &\theta_3 \sim N_+(\mu_{\theta_3}, \sigma^2_{\theta_3}),\ \theta_4 \sim LN(\mu_{\theta_4}, \sigma^2_{\theta_4}) \\ &\omega = 1/(1+\tau^2) \sim \mathsf{Beta}(a_\omega, b_\omega),\ a_\omega = 0.5,\ b_\omega = \exp(-\log(n)/2), \end{aligned}$$

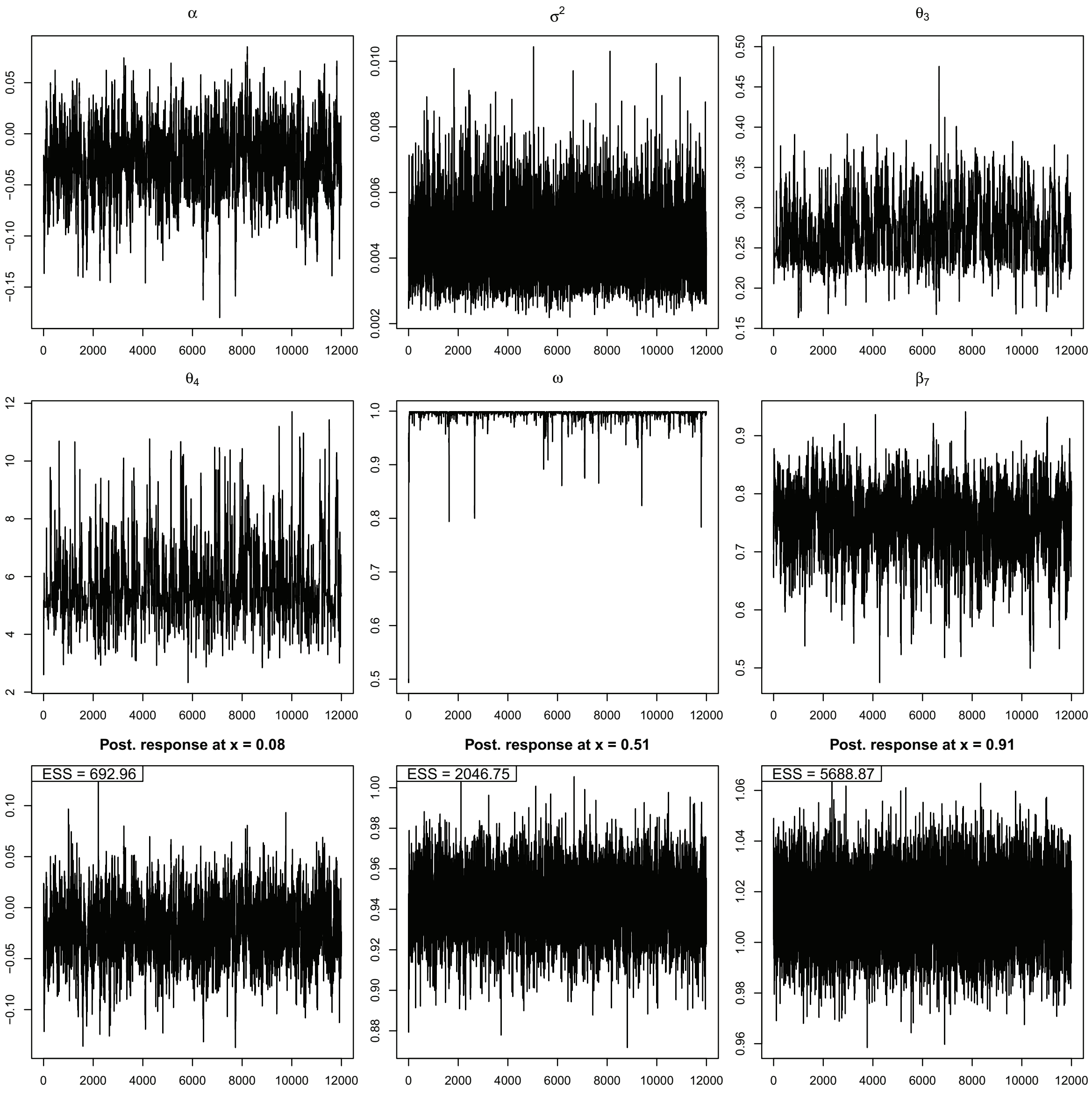


Figure 4: Traceplots of the NLFS(Hill) example fit in Figure 2, pane (a), correct subspace specification. The first 2000 samples were discarded as burn-in. Due to the correct subspace specification, there is strong shrinkage ($\omega = (1 + \tau^2)^{-1}$ close to 1). The effective sample size (ESS) was calculated based on the 10000 draws after discarding the first 2000 burn-in draws using the `coda` R-package [Plummer et al., 2006].

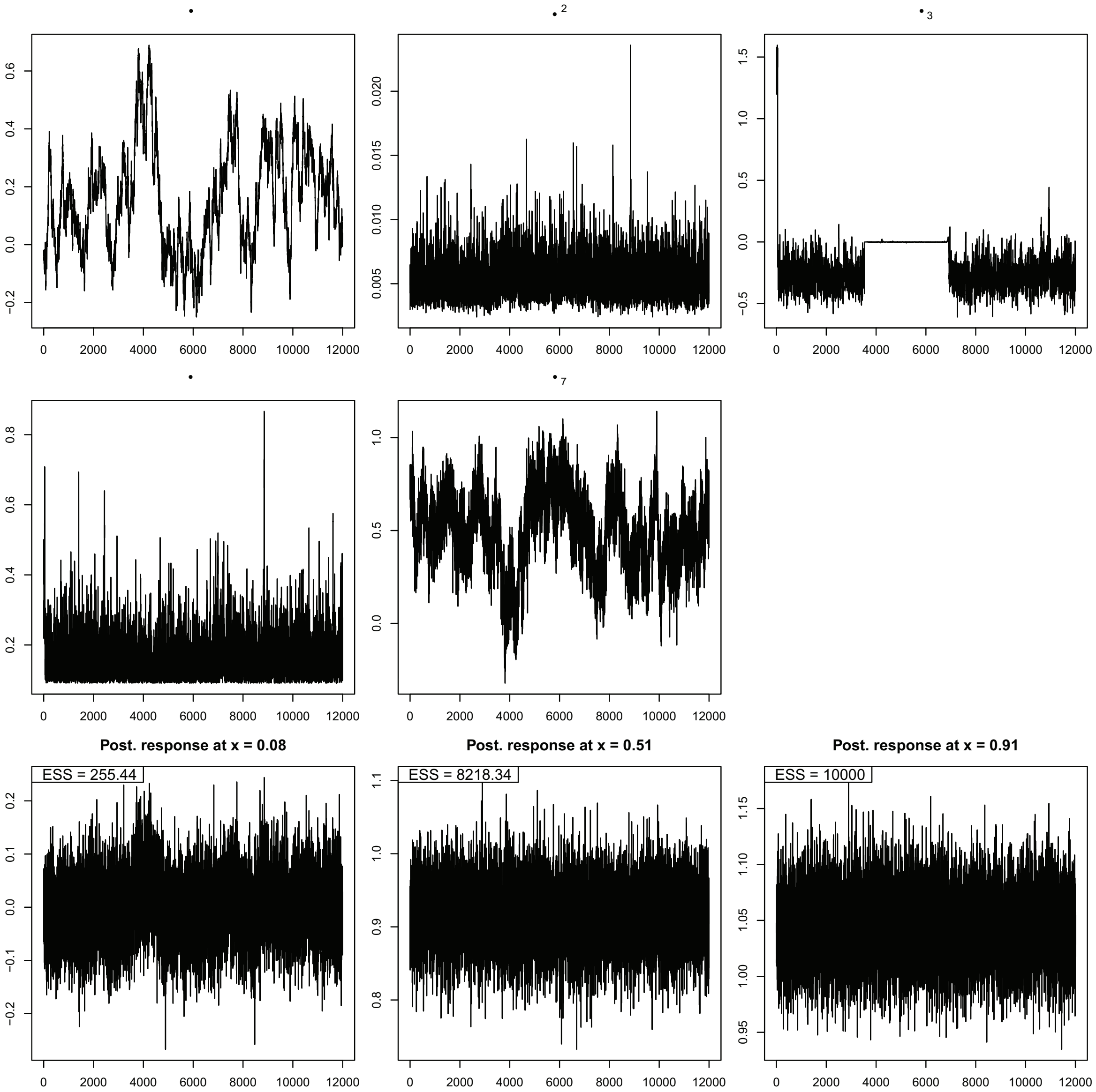


Figure 5: Traceplots of the NLFS(power) example fit in Figure 2, pane (b), subspace misspecification. The first 2000 samples were discarded as burn-in. Due to the subspace misspecification, there is little shrinkage ($\omega = (1 + \tau^2)^{-1}$ close to 0). Further, power exponent θ_3 is stuck at 0 for a few thousand draws, and θ_1, the intercept, seems highly correlated. Due to the misspecification and effectively no shrinkage, single parameters are not well identifiable and the whole response must be viewed to inspect convergence. The resulting response mixes well (bottom row). The effective sample size (ESS) was calculated based on the 10000 draws after discarding the first 2000 burn-in draws using the `coda` R-package [Plummer et al., 2006].

Algorithm 1 Non-linear functional shrinkage (NLFS)

```
Initialize: $\beta^{(1)}, \sigma^{2(1)}, \tau^{(1)}, \omega^{(1)}, \theta^{(1)}$
for $i : 2 \to B$ do
    Calculate $\dot{\mathsf{H}}_{\theta^{(i-1)}}$
    Sample $\beta^{(i)} \sim p(\beta|\cdot)$                              ▷ Conjugate
    Sample $\theta_1^{(i)} \sim p(\theta_1|\cdot)$                        ▷ Conjugate
    Sample $\sigma^{2(i)} \sim p(\sigma^2|\cdot)$                         ▷ Conjugate
    Sample $\omega^{(i)} \sim p(\omega|\cdot)$                            ▷ Slice Sampler
    Sample Non-linear $\theta^{(i)} \sim p(\theta|\cdot)$                 ▷ MH Sampler
end for
return All samples
```

the sampler is summarized in Algorithm 1.

Update β

We update β using the full conditional posterior

$$\beta|\sigma^2, \tau^2, \theta, y \sim N(\mu'_\beta, \Sigma'_\beta) \tag{14}$$

where $\Sigma'_\beta = \sigma^2(\Phi^\top\Phi + \tau^2\Phi^\top(I_n - P_\theta)\Phi)^{-1}$ and $\mu'_\beta = \sigma^{-2}\Sigma'_\beta\Phi'\tilde{y}$ and $\tilde{y} = y - 1_n\theta_1$.

Update θ_1

The intercept θ_1 is updated using

$$\theta_1|\beta, \sigma^2, y \sim N(\mu'_{\theta_1}, {\sigma^2_{\theta_1}}') \tag{15}$$

where ${\sigma^2_{\theta_1}}' = (\sigma^2\sigma^2_{\theta_1})(n\sigma^2_{\theta_1} + \sigma^2)$ and $\mu'_{\theta_1} = \sigma^{-2}{\sigma^2_{\theta_1}}'1_n^\top(y - \Phi\beta) + \mu_{\theta_1}/\sigma^2_{\theta_1}$.

Update σ^2

The noise variance σ^2 is updated by

$$\sigma^2|\beta, \tau^2, \theta \sim \text{IG}(a'_\sigma, b'_\sigma) \tag{16}$$

where $a'_\sigma = (n+k)/2 + a_\sigma$ and $b'_\sigma = 0.5(\text{RSS} + \tau^{-2}\beta^\top(\Phi^\top(I_n - P_\theta)\Phi)\beta) + b_\sigma$ where $RSS = ||y - (\theta_1 1_n + \Phi\beta)||_2^2$ is the residual sum of squares and $||.||_2$ is the Euclidean norm.

Update τ^2

We update τ^2 using a slice sampler [Neal, 2003] considering the posterior log likelihood

$$\begin{aligned} g(\tau) = \log(p(\tau|\beta, \sigma^2, \theta)) =& (-k/2 + b_\omega - 0.5)\log(\tau^2) && (17)\\ &+ (-a_\omega - b_\omega)\log(1 + \tau^2) && (18)\\ &+ \left(-\frac{1}{2\sigma^2}\beta^\top\Phi^\top(I_n - P_\theta)\Phi\beta\right)\tau^{-2}. && (19) \end{aligned}$$

Note that we use $-k/2$ and not $-(k-d_0)/2$ as in Shin et al. [2020] where d_0 is the rank of the (linear) projection matrix. We omit $-d_0$ as for the non-linear approach, there are no linear bases in $\dot{\mathsf{H}}_\theta$ that are in Φ (because there is no intercept in Φ) and hence the prior covariance matrix of β is of full rank k. For a current τ_0, calculate $v = g(\tau_0))$. Uniformly draw $\tilde{z} \sim U(0, \exp(v))$. For $z = \log(\tilde{z})$ define the slice $S_z = \{x : g(x) < g(z)\}$ and sample the next τ_1 uniformly from S_z. For computational ease, we restrict τ^2 to [0.001, 10].

The above sampling for the general $\omega \sim \text{Beta}(a_\omega, b_\omega)$ prior was primarily featured and labelled 'own slice' (OC) in Table 3. We also considered a standard horseshoe (HS) prior ($a = b = 0.5$). Details on its implementation are in Makalic and Schmidt [2015].

Update non-linear θ

One only has to update the non-linear parameters of θ, as P_θ only depends on the non-linear parameters. For the Hill model, the non-linear parameters are θ_3 and θ_4. We assume independence and separately update θ_3 and θ_4 using a Metropolis-Hastings sampler [Brooks et al., 2011] and explain the sampling for θ_3.

Perform the three sampling steps

1. Draw a candidate $\theta_3^{(1)}$ from a proposal distribution p_{prop}using $\theta_3^{(0)}$
2. Calculate the hastings ratio
$$\text{HR} = \frac{p(\theta_3^{(1)}|\cdot)p_{\text{prop}}(\theta_3^{(0)}|\theta_3^{(1)})}{p(\theta_3^{(0)}|\cdot)p_{\text{prop}}(\theta_3^{(1)}|\theta_3^{(0)})}.$$
3. Draw $u \sim \text{Unif}[0,1]$. If HR $> u$, accept $\theta_3^{(1)}$ as new draw. Otherwise, reject and consider $\theta_3^{(0)}$ as new draw.

For step (1), sample a new candidate $\theta_3^{(1)}$ from a proposal distribution, e.g. $N_+(\theta_3^{(0)}, \sigma^2_{\text{prop}})$ where σ^2_{prop} might be calculated as, e.g. the empirical variance of the latest 100 draws of θ_3, or simply as $\sigma^2_{\text{prop}} = \sigma^2_{\theta_3}$. To sample x from a truncated normal distribution with positive support, $x \sim N_+(\mu, \sigma^2)$, calculate $l = P(X < 0)$ where $X \sim N(\mu, \sigma^2)$. Sample $u \sim \text{Unif}[l, 1]$ and calculate $x = q_{\mu,\sigma^2}(u)$, the corresponding quantile.

For step (2), consider $\log(\text{HR})$ for computational stability:

$$\log(\text{HR}) = \log(p(\theta_3^{(1)}|\cdot)) - \log(p(\theta_3^{(0)}|\cdot)) + \log(p_{\text{prop}}(\theta_3^{(0)}|\theta_3^{(1)})) - \log(p_{\text{prop}}(\theta_3^{(1)}|\theta_3^{(0)})).$$

For the fully conditional log posterior $\log(p(\theta|\cdot))$, we can integrate out β to reduce the autocorrelation in the sampling. Since $p(\theta|\cdot) \propto p(y|\theta, \sigma^2, \tau^2)p(\theta)$ and

$$y|\theta, \sigma^2, \tau^2 \sim N(\mathbb{E}(\Phi\beta + \varepsilon), \text{Cov}(\Phi\beta + \varepsilon)),$$

we use $y|\theta, \sigma^2, \tau^2 \sim N(0, \Sigma_y)$ with $\Sigma_y = \sigma^2\tau^2\Phi(\Phi^\top(I - P_\theta)\Phi)^{-1}\Phi^\top + \sigma^2 I_n$.

References

M. A. Alvarez, D. Luengo, and N. D. Lawrence. Linear latent force models using gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2693–2705, 2013.

A. Brezger and W. J. Steiner. Monotonic regression based on bayesian p–splines. *Journal of Business & Economic Statistics*, 26(1):90–104, 2008. doi: 10.1198/073500107000000223. URL `https://doi.org/10.1198/073500107000000223`.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

D. Carl. *A practical guide to splines*. Springer, 2001.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

J. Chen, Z. Chen, C. Zhang, and C. Jeff Wu. Apik: Active physics-informed kriging model with partial differential equations. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):481–506, 2022.

A. Golightly and D. J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–820, 2011.

S. Goutelle, M. Maurin, F. Rougier, X. Barbaut, L. Bourguignon, M. Ducher, and P. Maire. The hill equation: a review of its capabilities in pharmacological modelling. *Fundamental & clinical pharmacology*, 22(6):633–648, 2008.

L. H. Gunn and D. B. Dunson. A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, 6(3):434–449, 2005.

A. V. Hill. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *The Journal of Physiology*, 40:iv–vii, 1910.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Y. Huang, D. Liu, and H. Wu. Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics*, 62(2):413–423, 2006.

T. W. Kelsey, L. Q. Li, R. T. Mitchell, A. Whelan, R. A. Anderson, and W. H. B. Wallace. A validated age-related normative model for male total testosterone shows increasing variance but no decline after age 40 years. *PloS one*, 9(10):e109346, 2014.

C. Köllmann, B. Bornkamp, and K. Ickstadt. Unimodal regression using bernstein–schoenberg splines and penalties. *Biometrics*, 70(4):783–793, 2014.

S. Lang and A. Brezger. Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004.

D. J. Lunn, N. Best, A. Thomas, J. Wakefield, and D. Spiegelhalter. Bayesian analysis of population pk/pd models: general concepts and software. *Journal of pharmacokinetics and pharmacodynamics*, 29:271–307, 2002.

E. Makalic and D. F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

D. Mesarovic, Mihajlo, S. Sreenath, and J. Keene. Search for organising principles: understanding in systems biology. *Systems biology*, 1(1):19–27, 2004.

M. C. Meyer. Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, 2(3):1013–1033, 2008. ISSN 19326157. URL `http://www.jstor.org/stable/30245118`.

M. C. Meyer, A. J. Hackstadt, and J. A. Hoeting. Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *Journal of Nonparametric Statistics*, 23(4):867–884, 2011.

T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

R. M. Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

M. Plummer, N. Best, K. Cowles, and K. Vines. Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006. URL `https://journal.r-project.org/archive/`.

G. A. Seber and C. J. Wild. Nonlinear regression. hoboken. *New Jersey: John Wiley & Sons*, 62 (63):1238, 2003.

M. Shin, A. Bhattacharya, and V. E. Johnson. Functional horseshoe priors for subspace shrinkage. *Journal of the American Statistical Association*, 115(532):1784–1797, 2020.

T. S. Shively, T. W. Sager, and S. G. Walker. A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(1):159–175, 2009.

T. S. Shively, S. G. Walker, and P. Damien. Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161(2):166–181, 2011.

P. Šimon. Considerations on the single-step kinetics approximation. *Journal of Thermal Analysis and Calorimetry*, 82(3):651–657, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

M. K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray. Identifying targets of multiple coregulating transcription factors from expression time-series by bayesian model comparison. *BMC systems biology*, 6:1–21, 2012.

M. W. Wheeler, D. B. Dunson, S. P. Pandalai, B. A. Baker, and A. H. Herring. Mechanistic hierarchical gaussian processes. *Journal of the American Statistical Association*, 109(507): 894–904, 2014.

M. W. Wheeler, D. B. Dunson, and A. H. Herring. Bayesian local extremum splines. *Biometrika*, 104(4):939–952, 2017.

P. Wiemann and T. Kneib. Adaptive shrinkage of smooth functional effects towards a predefined functional subspace. *arXiv preprint arXiv:2101.05630*, 2021.